



**This electronic thesis or dissertation has been
downloaded from Explore Bristol Research,
<http://research-information.bristol.ac.uk>**

Author:
Darley, Emily

Title:
Negating incrementally

The impact of contextual predictability on processing of affirmative and negative sentences

General rights

Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>. This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy

Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

NEGATING INCREMENTALLY: THE IMPACT OF
CONTEXTUAL PREDICTABILITY ON PROCESSING OF
AFFIRMATIVE AND NEGATIVE SENTENCES

Emily J. Darley

A dissertation submitted to the University of Bristol in accordance with the
requirements for award of the degree of Doctor of Philosophy in the School of
Psychological Science, Faculty of Life Sciences.

September 2018

40,779 words

Abstract

Language processing is normally rapid, incremental and driven by online prediction-making. However, the phenomenon of negation presents an interesting possible exception to this case. Although at least some utterances involving negation seem to induce effects on processing in which predictions cannot be made or their accuracy breaks down, evidence suggests that this is not the case when negations are presented in contexts providing adequate pragmatic support. This thesis presents a series of experiments using several methodologies (EEG, computer mouse-tracking, and eye-tracking) to test the idea that this effect of pragmatic felicity can be attributed to its association with predictability: that is, to investigate whether the predictability of later material in a sentence can influence the extent to which negation can be incorporated incrementally into the comprehender's interpretation of the partial sentence. This is achieved through the use of episodically-varying contexts, presented prior to the accompanying linguistic input, which manipulate the predictability of critical material in pragmatically licensed sentences. The findings lead to the overall conclusion that even pragmatically licensed negations can incur more processing costs and result in the generation of more inaccurate predictions than equivalent affirmatives. Furthermore, the most reliable strands of evidence (from a sentence completion mouse-tracking task) suggest that, in the type of paradigm in which prediction is manipulated using episodic contexts, reducing predictability detrimentally affects affirmatives to a greater extent than negations. This may indicate that accurate prediction-making is relatively difficult, even in the easiest cases, for negative sentences when (1) predictions must be formulated on the basis of episodic rather than long-term semantic associations and (2) the combination of sentence structure and context mean that there is a clash between the concepts activated by association with local components of the sentence and those relevant to its global interpretation.

Acknowledgements

This PhD was carried out under the supervision of Nina Kazanina and Chris Kent, both of whom have contributed enormously to the ideas, experimental designs, and data analyses, as well as providing extremely valuable input on the write-up of all chapters. I have tested their support and patience to what ought to have been the limit, and they have continued to give their time and expertise with unreasonable generosity. Thank you so much. Although Nina and Chris's indefatigable attention to detail has prevented countless mistakes of mine from sneaking through, I am of course responsible for those that remain.

I would also like to thank various other members and former members of the School for their help and contributions. My annual assessors, Josie Briscoe and Steve Lewandowsky, provided excellent sounding boards for the general direction of my research every year, and reassurance that my work was on track. Useful guidance in the lab was provided by Phil Collard, George Stothart and Jen Todd Jones, and I have exchanged very useful discussions on analysis, interpretation and statistical techniques with many people, including Emily Crowe, Michele Gubian, Niall Taylor and Rob Udale. Many undergraduate research assistants have helped out with various activities: thanks to Stella Becci, Mischa Dhar, Hugo Hammond, Julie Lee, Jamie Mcevoy, Shanaz Pottinger, Alesi Rowland, Zoe Travers and Olivia Winton for assistance with data collection; to Holly Eager and Barry O'Mahony for assistance with auditory and visual stimulus creation; and to Katherine Tapp for assistance with data cleaning. Finally, the community of postgraduate students in the School of Experimental Psychology has been wonderfully warm and supportive: a fantastic environment in which to exchange ideas and frustrations.

On a visit to the University of Maryland in 2015, I was warmly welcomed by what felt like the entire Department of Linguistics. Thanks to Colin Phillips for supervising my work during this visit and for many useful discussions; I also enjoyed interesting and helpful conversations with Allyson Ettinger, Julie Gerard, Shota Momma, Lara Ehrenhofer and Zoe Schlueter, among others. Thanks to Julia Buffinton and Jon Burnsky for helping to set up and run the eye-tracking study, and to Jon along with Hanna Muller for taking charge of that project to spin it off in further interesting directions.

I also gratefully acknowledge several funding sources. A 1+3 studentship from the Economic and Social Research Council made my PhD financially possible. My

time at the University of Maryland was generously funded jointly by an Overseas Institutional Visit award from the Southwest Doctoral Training Centre and an International Graduate Research Fellowship from the University of Maryland. Contributions along the way from Guarantors of Brain and the Bristol Alumni Foundation enabled me to attend several conferences and workshops which enriched the experience and allowed me to obtain valuable feedback from others in the field.

The word *essential* doesn't even begin to describe the role played in completing this work by the support of my family and friends, who have rallied around me in impossibly difficult times. Thank you for the reasons to persist — to my parents, Zoe, all of the Russells, and the innumerable friends who have stepped up quietly and patiently and without being asked, to bring forms of help that I didn't even realise I needed. The Girtonians of the TFR group deserve a special mention: thank you for the outlet for both silliness and realism in equal measure. Your friendship means so much to me.

My partner Nick's support and belief in me made my starting a PhD possible, and his loss made my completing it next to impossible. I hope he'd be proud that I made it to the finishing line eventually, although the finished product is poorer not only for the devastating blow to my working capacity and quality of life in general, but also for the absence of my partner in enthusiastic and wide-ranging idea-tennis over the dinner table. My thoughts always used to come back to me with new embellishments, looking shinier and more promising. The best I can do in lieu of the many further contributions he would have made is to take this opportunity to add something he'd appreciate: GNU Nick Russell.¹

¹See Pratchett, T. (2004). *Going Postal*. London: Doubleday. Hardback edition pp. 98–99.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED:

DATE:

Contents

1	General Introduction	19
1.1	Incrementality and prediction-making	20
1.2	Background on negation	21
1.2.1	Approaches from philosophy and semantics	21
1.2.2	Psycholinguistic work so far	26
1.3	Introduction to experimental work	32
2	EEG Study	34
2.1	Introduction	34
2.1.1	Electroencephalography in language research	34
2.1.2	Using the N400 to investigate incrementality and negation	35
2.1.3	The role of pragmatic licensing and predictability	39
2.1.4	The present study	40
2.1.5	Design and hypotheses	43
2.2	Methods	44
2.2.1	Materials	44
2.2.2	Participants	47
2.2.3	Procedure	48
2.3	Results	50
2.3.1	Behavioural response accuracy	50
2.3.2	Event-related potentials	51
2.3.3	Memory test	57
2.4	Discussion	58
3	Mousetracking Experiment 1: Truth Value Judgement	64
3.1	Introduction	64
3.1.1	Negation and predictability	64

3.1.2	The mouse-tracking methodology	66
3.1.3	Design and hypotheses	67
3.2	Methods	69
3.2.1	Materials	69
3.2.2	Procedure	72
3.2.3	Participants	73
3.3	Results	74
3.3.1	Data preparation and analysis	74
3.3.2	Response accuracy	78
3.3.3	Response initiation time	82
3.3.4	Response completion time	86
3.3.5	Trajectory shapes	90
3.3.6	Clustering of trajectories	93
3.3.7	Memory test	97
3.4	Discussion	97
4	Mousetracking Experiments 2 and 3: Sentence Completion	102
4.1	Introduction	102
4.2	Methods	103
4.2.1	Materials	103
4.2.2	Procedure	106
4.2.3	Participants	107
4.3	Results	108
4.3.1	Data preparation and analysis	108
4.3.2	Statistical modelling of effects	108
4.3.3	Experiment 2	109
4.3.4	Experiment 3	122
4.4	Discussion	135
4.4.1	Was there a floor effect?	138
4.4.2	Anti-prediction	138
4.4.3	Interpreting the findings	140
4.4.4	Conclusions	145
5	Eyetracking Study	148
5.1	Introduction	148
5.1.1	Background	148
5.1.2	The present study	151

5.2	Methods	152
5.2.1	Materials	152
5.2.2	Participants	154
5.2.3	Procedure	154
5.3	Results	156
5.3.1	Comprehension response accuracy	156
5.3.2	Eye movements	156
5.4	Discussion	162
6	General Discussion	166
6.1	Summary of motivation	166
6.2	Summary of findings	167
6.3	Situating the findings relative to other work	169
6.4	Global and local processing	172
6.5	Limitations and generalisability	175
6.6	Future work	176
6.7	Conclusion	177
	Bibliography	178

List of Tables

2.1	EEG experiment linguistic stimuli	45
2.2	EEG experiment simple effects on response accuracy	52
2.3	EEG experiment simple effects on N400 amplitude	56
2.4	EEG experiment memory test results	58
3.1	Mouse-tracking Experiment 1 example stimulus sets	71
3.2	Mouse-tracking Experiment 1 simple effects on response accuracy	81
3.3	Mouse-tracking Experiment 1 simple effects on initiation time . .	85
3.4	Mouse-tracking Experiment 1 simple effects on response time . .	89
3.5	Mouse-tracking Experiment 1 simple effects on trajectory cluster	96
4.1	Mouse-tracking Experiments 2 and 3 example stimulus sets	105
4.2	Mouse-tracking Experiment 2 simple effects on response accuracy	110
4.3	Mouse-tracking Experiment 2 simple effects on initiation time . .	113
4.4	Mouse-tracking Experiment 2 simple effects on response time . .	115
4.5	Mouse-tracking Experiment 2 simple effects on trajectory cluster	121
4.6	Mouse-tracking Experiment 3 simple effects on response accuracy	123
4.7	Mouse-tracking Experiment 3 simple effects on initiation time . .	126
4.8	Mouse-tracking Experiment 3 simple effects on response time . .	128
4.9	Mouse-tracking Experiment 3 simple effects on trajectory cluster	133
4.10	A summary of the main findings across all three mouse-tracking experiments	137
5.1	Eye-tracking experiment simple effects on proportion of looks during the critical window	160
5.2	Further eye-tracking experimental conditions	164

List of Figures

1.1	Square of opposition	22
2.1	EEG experiment visual context stimulus	45
2.2	Electrode layout in EEG experiment	48
2.3	EEG experiment response accuracy violin plots	50
2.4	Mean timecourse of EEG in the window of interest	54
2.5	EEG experiment N400 results	57
2.6	Re-visualisation of mean timecourse of EEG	60
3.1	Mouse-tracking Experiment 1 example trial	73
3.2	Area under the curve and maximum deviation in mouse-tracking	75
3.3	Mouse-tracking Experiment 1 response accuracy violin plots	79
3.4	Mouse-tracking Experiment 1 response accuracy effects	80
3.5	Mouse-tracking Experiment 1 initiation time distribution and violin plots	83
3.6	Mouse-tracking Experiment 1 initiation time	84
3.7	Mouse-tracking Experiment 1 response time distribution and violin plots	87
3.8	Mouse-tracking Experiment 1 response time effects	88
3.9	All response trajectories in mouse-tracking Experiment 1	90
3.10	Mouse-tracking Experiment 1 heat map of cursor locations	91
3.11	Mouse-tracking Experiment 1 area under the curve and maximum deviation distributions	92
3.12	Mouse-tracking Experiment 1 maximum deviation across conditions	93
3.13	Mouse-tracking Experiment 1 individual participant trajectories	94
3.14	Mouse-tracking Experiment 1 trajectory clusters	94
3.15	Mouse-tracking Experiment 1 effects on cluster trajectory	95

4.1	Mouse-tracking Experiments 2 and 3 example trial	106
4.2	Mouse-tracking Experiment 2 response accuracy violin plots . . .	110
4.3	Mouse-tracking Experiment 2 response accuracy effects	111
4.4	Mouse-tracking Experiment 2 initiation time distribution and violin plots	112
4.5	Mouse-tracking Experiment 2 initiation time effects	112
4.6	Mouse-tracking Experiment 2 response time distribution and violin plots	114
4.7	Mouse-tracking Experiment 2 response time effects	114
4.8	All response trajectories in mouse-tracking Experiment 2	116
4.9	Mouse-tracking Experiment 2 heat map of cursor locations	117
4.10	Mouse-tracking Experiment 2 area under the curve and maximum deviation distributions	119
4.11	Mouse-tracking Experiment 2 maximum deviation across conditions	119
4.12	Mouse-tracking Experiment 2 individual participant trajectories .	120
4.13	Mouse-tracking Experiment 2 trajectory clusters	120
4.14	Mouse-tracking Experiment 2 effects on cluster trajectory	121
4.15	Mouse-tracking Experiment 3 response accuracy violin plots . . .	123
4.16	Mouse-tracking Experiment 3 response accuracy effects	124
4.17	Mouse-tracking Experiment 3 initiation time distribution and violin plots	125
4.18	Mouse-tracking Experiment 3 initiation time effects	126
4.19	Mouse-tracking Experiment 3 response time effects	127
4.20	Mouse-tracking Experiment 3 response time distributions and violin plots	128
4.21	All response trajectories in mouse-tracking Experiment 3	129
4.22	Mouse-tracking Experiment 3 heat map of cursor locations	130
4.23	Mouse-tracking Experiment 3 area under the curve and maximum deviation distributions	131
4.24	Mouse-tracking Experiment 3 maximum deviation across conditions	131
4.25	Mouse-tracking Experiment 3 individual participant trajectories .	132
4.26	Mouse-tracking Experiment 3 trajectory clusters	133
4.27	Mouse-tracking Experiment 3 effects on trajectory cluster	134
4.28	Cluster allocations for varying-predictability and constant-predictability trials in Experiments 2 and 3	142

5.1	Example pair of eye-tracking experiment image stimuli	153
5.2	Eye-tracking experiment example trial	155
5.3	Eye-tracking experiment: time course of eye movements during an average trial in each condition	158
5.4	Eye-tracking experiment effects on proportion of looks	161

Chapter 1

General Introduction

I have a vague early childhood memory of one of my parents occasionally handing me a drink accompanied by a playful injunction to “keep it in the cup!”. Presumably, their theory was that the negative instruction “don’t spill it!” would have involved mentioning the very event that it sought to avoid, thereby increasing the already all-too-high chances of an accident when putting a toddler’s clumsy hands and suggestible mind in charge of anything. This idea may not have been far off the mark in terms of how negations are processed, even by adult comprehenders.

As outlined below, incoming linguistic input is generally interpreted incrementally, as it becomes available, but negation seems to be an exception to this rule, at least under some circumstances. This phenomenon makes negation interesting in at least two respects: first, in its own right, as a kind of special case in terms of how processing must be handled; and second, as a “failure case” for incrementality. Investigating the circumstances under which a system breaks down is often a good way of understanding some aspect of how the system works when it does.

The remainder of this introductory chapter is organised as follows. First, the evidence that linguistic processing normally does proceed incrementally is briefly reviewed. Second, a survey of views on the phenomenon of negation is presented; this consists of a set of theoretical and philosophical approaches, followed by a review of the psycholinguistic literature that is more directly pertinent to the experiments described in this thesis. Finally, an overview of the experimental chapters and their general motivation is provided.

1.1 Incrementality and prediction-making

Speech presents a notoriously difficult perception and processing problem. Listeners are confronted with an unbroken stream of auditory input, continuously unfolding over time and consisting of acoustic signals that do not map directly onto linguistic units, but depend on characteristics of the speaker, environmental surroundings, and context of a particular segment (see Nygaard & Pisoni, 1995, for a review of these challenges). For these reasons, reliance solely on bottom-up processing of the input would render the task of making sense of it impossible; instead, comprehenders must draw on their knowledge of the language, speaker, discourse context, pragmatic norms, and so on to engage in top-down processing, in which partial representations of the utterance so far are continuously and rapidly updated, and predictions formulated against which upcoming material can be evaluated: this view is presented, for example, by Altmann and Mirković (2009) and by Rayner and Clifton (2009). Pickering and Garrod (2007) argue that this process is supported in particular by the production system, and Sturt and Lombardo (2005) show that the continuously updated representation takes the form of a connected syntactic structure — although such structures may not always be consistent with the full input, especially when processing demands are high, and must sometimes be reformulated when there is a clash between what is apparent locally and the globally correct structure (e.g. Ferreira, Bailey, & Ferraro, 2002).

There is extensive evidence of many specific cases in which comprehenders engage in active and incremental prediction. For instance, a verb can constrain comprehenders' predictions about likely upcoming arguments based on semantics, such as the knowledge that the verb *eat* is likely to be followed by an edible object such as *cake* (Altmann & Kamide, 1999). Similarly, arguments preceding a verb allow comprehenders to constrain their predictions about the semantic and syntactic characteristics of the verb phrase (Kamide, Altmann, & Haywood, 2003; Kazanina, 2017). Such predictions are not simply reliant on the lexical associations of preceding material, but are sensitive to the specific thematic roles of arguments (Chow, Smith, Lau, & Phillips, 2016) and to pre-existing knowledge about how events occur (Kim, Oines, & Sikos, 2016). However, the full extent and nature of these prediction-making processes is still the subject of investigation and debate. For instance, a finding that predictions about the semantics of an upcoming word can generate predictions relating to the phonological form of a preceding article (DeLong, Urbach, & Kutas, 2005) has recently been challenged by a large-scale

study in which it was not replicated (Nieuwland et al., 2018).

Negation presents a case in which comprehenders' predictions often appear to fail to take into account all the available information immediately: specifically, the presence of the negating element itself. The details of this phenomenon are discussed below, following a survey of the nature of negation itself.

1.2 Background on negation

1.2.1 Approaches from philosophy and semantics

A discussion of some of the historical and philosophical approaches to studying negation may help to illuminate why its behaviour in language might exhibit some properties worth studying: in particular, why we might a priori expect the presence of negation to have an impact on the online, incremental processing of sentences, and what form this impact might take.

Negation has troubled philosophers, logicians and linguists, all in slightly different ways, over the centuries. The crux of the problem it presents is that it seems simultaneously both primitive or basic — you can assert either that something *is*, or that it is *not* — and also slippery, abstract and complex — what does it really mean to assert that something lacks a particular property? Is every negation simply the reversal of a corresponding affirmative, or is there more to it than that? Can any given negation be couched as an affirmative equivalent, and if so, why do we bother with them if we could simply express any proposition using the far less perplexing affirmative? Negation seems to be a universal of natural human languages (Greenberg, 1966), and at least some aspects of negation appear developmentally early, although others prove difficult for young children to grasp (Nordmeyer & Frank, 2014; Wode, 1977). Negating particles like *not* seem fundamental to their languages and are extremely frequent, and yet they are unusually subject to clines of lexicalisation or fossilisation, in which their semantic force is gradually eroded and they are eventually replaced, often by a former intensifier — a process known as Jespersen's Cycle following the original observation by Jespersen (1917). This set of puzzles leads to a variety of approaches and analyses, some of which provide important background for understanding online comprehension of negation.

Horn (1989) provides a history of philosophical thinking on the topic, beginning with Plato, who articulates the concept of negation in ontological terms, and Aristotle, who supplies an early framework for negation in language and logic.

Early approaches concentrated on categorising negation into various sub-types of relationship: following Horn's (1989) examples, Aristotle's taxonomy includes contrariety (e.g. the relationship between *good* and *bad*), privation (*blind* / *sighted*), contradiction (*he sits* / *he does not sit*), and correlations between two relative terms (*double* / *half*), while the Stoics recognise three varieties: denial (*no one is walking*), privation (*this man is unkind*) and negation (*it is not the case that it is day*). The descendants of such taxonomic systems are still seen today in the "square of opposition" (e.g., Sullivan, 1967) shown in Figure 1.1, which presents the relationships between quantifying propositions. Aristotle also anticipates important foundations of the logic of opposition which are still debated, including the law of contradiction (it is impossible to be both P and $\neg P$ simultaneously) and the law of the excluded middle (either P or $\neg P$ must be true in every case).

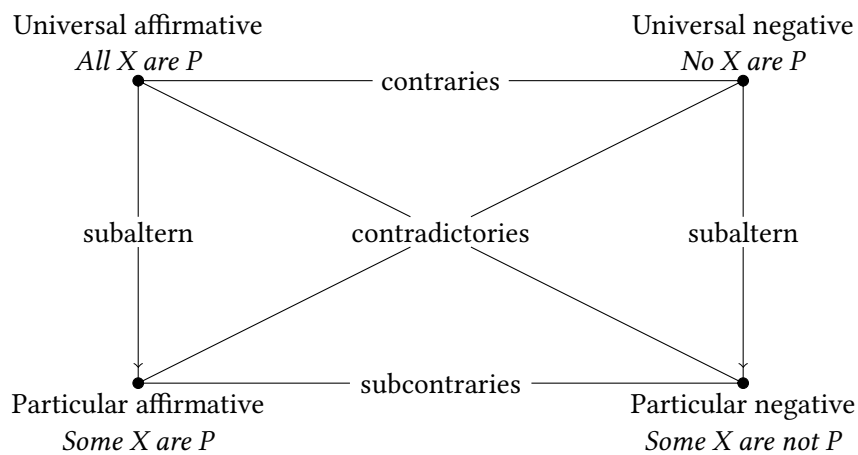


Figure 1.1: Square of opposition

From the perspective of a philosopher or logician, understanding and formalising negation should involve either defining it rigorously, or demonstrating that it is not a necessary component of a system by reducing it to more basic terms, usually an affirmation plus some concept of falsehood. (The opposite approach is also possible: for instance, Löbner (2000) proposes a definition of falsity in terms of the truth value of syntactic negations.) It is important to distinguish this problem from the problem of defining linguistic negation by simply pointing to negating words like *not* in the linguistic form of a proposition. (Even this apparently more straightforward classification is not without its grey areas. For instance, many semantically rather than syntactically negative words license negative polarity items (NPIs): contrast *He refused to come and visit ever again* with **He agreed to come and*

visit ever again. This grammatical effect of a semantic property of the verb *refuse* shows that there is more to linguistic negation than whether a sentence contains one of the obviously negating words *not*, *never* etc.) Philosophically speaking, the form a particular proposition happens to be expressed in is less interesting than whether there is some intrinsic asymmetry, regardless of particular linguistic form, between a negative and an affirmative, where affirmation is basic, simple and necessary, and negation is a layer of complexity built on top of this.

The “no” answer to this question is that of the symmetricalists, including Frege (1960) and Geach (1965) as well as Plato and Aristotle, whereas the asymmetricalist camp propose to “explain away” negation by formulating every negative proposition in terms of a corresponding affirmative. One approach to achieving this is a truth-conditional semantic one: the notion that the negation of a statement *P* is the statement that is true when *P* is false. Ayer (1952), also taking a symmetricalist view, reviews this approach and discusses the problems it presents: in order to check the relationship between an affirmative and a negative under this formalisation, a way of verifying either negative predicates or negative facts is needed. For example, to confirm that *The sky is not red* is the negation of *The sky is red*, it must be verified that the former is true whenever the latter is false, which in turn requires either defining a negative predicate *not red*, or defining a state of affairs where the negative fact that the sky is not red obtains, both of which prove epistemologically problematic. Ayer’s proposal is that rather than resorting to truth conditions and operations on affirmatives to identify negations, a distinction can be made between the two in that where a corresponding affirmative and negative pair can be identified, the negation will be less informative than the affirmation: for instance, *The sky is red* is much more informative than *The sky is not red*, which leaves a practically infinite variety of possible colours for the sky. This is a stance which puts negation and affirmation on an equal footing in the sense of distinguishing them without resorting to defining any given negation as some transformation of an affirmation. Another symmetricalist stance, rejectivism, is the notion that the primitive concept that needs to be handled in an analysis of language is not the negation operator, but the more pragmatically based attitude of rejection or speech act of denial: that is, negation should be explained in terms of these concepts, rather than vice versa. This case is laid out, for example, by Humberstone (2000) and Ripley (2011); there is further discussion of its application to psycholinguistic questions below.

In formalising how negation is interpreted, some modern approaches make use

of multi-valued logic, in which there is a third semantic value indicating neither TRUE nor FALSE, but an indeterminate, vague, or somewhere-in-between value. This allows for the distinction of three possible types of negation: choice negation (or strong negation), where $\neg P$ is defined as having the opposite truth-value to P ; exclusion negation (or weak negation), where $\neg P$ is defined as indicating “ P has some value other than TRUE”; and intuitionistic negation (or Gödel negation), where $\neg P$ is defined as indicating “ P has the value FALSE”; see, for example, Alxatib and Pelletier (2011) for discussion and experimental support for “vagueness” in speakers’ interpretation of negations, Shramko (2005) for an outline of the treatment and use of negation in “falsification logic”, a type of intuitionistic logic, and Wansing (2016) for another type of bi-intuitionistic logic, which combines intuitionistic verification and falsification in order to distinguish between support of truth and support of falsity. This approach therefore allows for the existence of affirmation-negation pairs that do not comply with the law of the excluded middle, permitting some fuzzy ground in which, for instance, speakers might regard it as correct to predicate both *tall* and *not tall* of a person of average height.

Another important problem in the interpretation of negation relates to presupposition, or failure to refer. This is classically presented — originally by Russell (1905) — as the question of whether the sentence *The present King of France is not bald* is true or false, or has some other truth value, given that there is no such entity as the King of France. Under an exclusion negation analysis, the sentence presupposes the existence of the predicate’s argument and predicates of it that it has the property of being not bald; whereas under a choice negation analysis, the sentence has no presupposition but simply denies the state of affairs that the King of France is bald: its truth conditions might be met by his failing to exist and therefore failing to have, among many other properties, the property of baldness. (A related problem is the category error problem, in which a proposition such as *Justice is not purple* denies the predication of something of an argument to which the predicate is not applicable.)

Aside from the handling of the truth conditional analysis problems presented by this presupposition problem, the choice-exclusion ambiguity can be treated under a syntactic analysis as one of scope, in which negation involves a covert syntactic transformation dictating the domain (in this case, either the predicate *bald* or the full proposition) over which the negation operates (Chomsky, 1970). The listener must resolve the ambiguity based on pragmatic considerations, which could range from contextual knowledge to intonational information or explicit cancellation of

the available presupposition by the speaker: *The King of France isn't bald; there is no King of France!* A semantic treatment of the issue is the Atlas-Kempson thesis, reviewed by Atlas (1977), which proposes that this type of negation is not necessarily ambiguous, but may be semantically “general, non-specific, or vague”. That is, although the speaker might intend either the choice or the exclusion interpretation in particular, they might also be in a state of equivocation between the two interpretations and intend to communicate the possibility of either to the listener.

The above set of problems, questions and approaches to analysis relating to negation range from the highly abstract and philosophical to those whose relevance to the interpretation of negation in natural language is clearer. Given the thorny collection of such issues that negation presents, it is worth returning to the question of why natural languages (seemingly without exception) deploy it at all. Would it not be easier to find a corresponding affirmative for every statement we wanted to make and use that instead? Price (1990) offers a rejectivist answer based on pragmatic rather than syntactic or semantic considerations — a reasonable approach given that ultimately humans interpret communication on a pragmatic level. He argues that negation has an evolutionarily crucial function, namely the ability to highlight and draw attention to incompatibilities between perspectives by explicitly rejecting an interlocutor's statement or a shared presupposition. A similar stance is formalised by Berto (2015) with a model of negation as a modal operator, in a possible worlds semantics framework, with a meaning based on the core concept of (in)compatibility.

Along with other evidence from approaches to interpreting negation, Price's argument for the importance not just of logical meaning but of directing the comprehender's attention to an implication of the speaker's meaning underscores the role of pragmatic context in understanding how negations are processed by comprehenders. The importance of this factor has become increasingly apparent as experimental and psycholinguistic work on negation has developed. For humans, context matters for logical operations: adjusting the pragmatic context of a task, even though its logical contents are unchanged, can have a dramatic effect on people's performance (Cheng & Holyoak, 1985; Cosmides, 1989; Manktelow & Over, 1991). As the next section will demonstrate by examining the psycholinguistics of negation and how it is processed online, negation seems to be no exception to this rule.

1.2.2 Psycholinguistic work so far

Early work in the psycholinguistics of negation focused on the comprehension of full propositions, demonstrating that at least superficially, there appeared to be an asymmetry in the processing of negations and affirmations. For example, Wason (1959, 1961) presented affirmative and negative statements with missing information, asking participants to fill in the required information to render them either true or false, and found that they were quicker to do so for affirmative sentences, regardless of target truth value, even though the task for false negations was effectively equivalent to the task for true affirmatives. Polarity also interacted with truth value, a finding replicated in a similar experiment by Gough (1965), who concluded that any hypothesis based on comprehenders performing a covert transformation of a negation into a “kernel” affirmative therefore could not fully account for this asymmetry. Wason (1961) proposed three possible explanations: that people find “positive information” to be more valuable, as it is more specific and informative, and therefore respond more readily to it; that “negative information”, being more abstract and therefore more distant from perceptual experience, is harder to access; and that negative terms may carry emotional connotations that disrupt the processing of expressions containing them.

Wason and Jones (1963) tested this last hypothesis with an experiment in which neutral nonsense words were used in place of the English word *not* to tag statements as affirmatives or negatives. Participants who reported mentally converting the nonsense word to a linguistic equivalent showed more similar response patterns to a control group tested on normal English sentences, compared to those who reported making their decisions based on the form of the nonsense word itself, suggesting that connotations of *not* might indeed partially account for the asymmetry. However, there were also early indications that this asymmetry might also be at least partially dependent on contextual factors: for instance, Johnson-Laird and Tridgell (1972) asked participants to draw inferences from pairs of premises, varying the forms of these premises. They found that when the first one provided a context in which a negation would be more helpful than an affirmative in drawing attention to an available inference (for example: “Either John is intelligent or he is rich. John is not rich”), people were quicker to respond than if an equivalent affirmative was presented instead.

To explore the details of what types of negation affect processing in this way, Just and Carpenter (1971) compared quantified negations in three forms, which they referred to as explicit syntactic negations (e.g. “None of the dots are red”),

implicit syntactic negations (e.g. “Few of the dots are red”), and semantic negations (e.g. “A minority of the dots are red”). Participants’ speed in making truth value judgments of these sentences with respect to a display were most different from comparable affirmatives in the case of semantic negations; the authors claimed that this was because syntactic negation focused their attention on the larger subset of items in the display, whereas semantic negation focused their attention on the smaller subset, which made the sentence more difficult to verify. Instructing participants explicitly on which type of encoding to use when looking at the display was successful in modulating this effect, showing that participants’ frame of attention and encoding was critical for how they processed negations. Thus, the process of shifting attention from a negated state of affairs to an actual state of affairs is what incurs extra costs.

More recently, research has examined the possible mechanisms underlying the processing and representation of negations. This work is motivated not only by hypotheses fundamentally distinguishing negations from affirmatives, but also by some lines of evidence that negation may present a case where fully incremental processing of language is not possible. Because of the challenges involved in speech perception, as mentioned above, it is well recognised that top-down processing, taking into account the comprehender’s pre-existing and updating knowledge of the context, speaker, current discourse referents, and so on, forms a major part of this task. Example demonstrations of such top-down processing and prediction include Sedivy, Tanenhaus, Chambers, and Carlson (1999) on the use of contextual information to anticipate relevant contrastive interpretations of adjectives, Kutas and Federmeier (2000) on the application of semantic knowledge to pre-activation of relevant concepts, and Hagoort and van Berkum (2007) on the parallel integration of information from multiple levels of interpretation during processing. As a component of this mode of processing, and as evidenced in these studies, comprehenders must constantly update their interpretations of partial utterances, actively forming partial representations and refining predictions that will help them to integrate upcoming material.

Nevertheless, some components of language lend themselves more readily than others to immediate integration into mental representations. Certain elements may require more time, more processing steps, or perhaps more information from still-unfolding parts of the utterance, before they can be fully integrated; these may include certain quantifiers (Huang & Snedeker, 2011; Urbach & Kutas, 2010) and scalar implicatures. The latter were investigated using a computer mouse-tracking

methodology by Tomlinson, Bailey, and Bott (2013), who presented participants with sentences of the form “Some elephants are mammals”. This has the literal and immediately available meaning “At least some elephants are mammals”, but also carries the implicature “Not all elephants are mammals”; the authors demonstrated that this appears to be derived only at a subsequent stage of processing. This indicates that the full suite of possible meanings of *some* is not equally active at the point of its incremental interpretation, and in fact the most relevant meaning is only brought to bear after the full proposition has been processed. Furthermore, sometimes there may be a clash between expectations suggested by different levels of analysis of the information encountered so far. Negation is a prime example of this type of clash, and this presents a possible explanation for the findings discussed above attributing processing difficulties to negation.

One possible factor in producing these difficulties is that the presence of a negating element such as *not* tends to generate situations in which activating concepts semantically related to or commonly collocated with the material so far directly opposes a prediction-making strategy that takes into account an incomplete semantic representation of the utterance and the Gricean assumption that the speaker will produce a true sentence. This may explain the finding by Fischler, Bloom, Childers, Roucos, and Perry (1983) that false sentences like “A robin is not a bird” elicit a reduced N400 ERP component (generally taken to indicate conformance with semantic expectations or ease of integration) in comparison to true sentences like “A robin is not a tree”. However, it is not always the case that semantic activation spreads indiscriminately of syntactic and semantic roles: predictions for upcoming verbs are sensitive to which preceding material is and is not an argument of the verb (Chow et al., 2016); and in the case of negation, concepts are activated to a lesser extent when negated (MacDonald & Just, 1989). This effect is also sensitive to whether or not the negation specifically indicates absence of an entity from the described situation, as in “Sam wished / was relieved that Laura was not wearing her pink dress” (Kaup, 2001; Kaup & Zwaan, 2003).

Another reason for asymmetric processing of the contents of affirmatives and negations may be that negations activate associations that are incongruent with the message of the full proposition. There are conflicting findings on this. MacDonald and Just (1989) used a probe task following a sentence in which either the probe or another noun was negated (e.g., “Elizabeth bakes some bread but no cookies”) and found that negation did suppress activation (as measured by response times to the probe) specifically for the noun with which it was associated. However,

the results did not follow this pattern when the probe was not the noun itself but an associate of the negated noun (“butter”/“cookies”). It is unclear whether this was due to small effects of spreading activation overall, or because negation failed to suppress associated concepts even though activation of the negated concept itself was suppressed. Mayo, Schul, and Burnstein (2004) found that the negations did activate concepts related to the negated concept: participants were slower to reject incongruent probes like “Tom’s clothes are folded neatly in his closet” after a negation “Tom is not a tidy person” than to reject incongruent probes after affirmatives. However, this was dependent on the type of negation presented: negations of predicates that could be conceptualised in a bipolar fashion, such as “not *tidy*” (i.e., *messy*), were more likely to be processed as integrated units (a “fusion” model of processing) than those for which this was not necessarily the case, such as *creative*. In addition to such effects on processing and initially responding to information presented in a negated form, research on the storage and access of such information has shown that this is sometimes difficult: encountering negated information may make people more likely to believe this information, even though they have been exposed to it in negated form (Gilbert, Krull, & Malone, 1990; Mayo, Schul, & Rosenthal, 2014).

Some researchers take the view that the information provided by negative statements is constructed by a kind of two-step perceptual simulation, in which the world described by the negated material is simulated and this information used to produce a simulation of the full proposition. This is a generalisation of the notion of representation through the use of possible world models, or “situation models”, in which perceptual aspects of a described world are simulated (e.g. Zwaan, Stanfield, & Yaxley, 2002) and deductions can be made by comparing and discarding possible models of the world (e.g. Barres & Johnson-Laird, 2003; Khemlani, Orenes, & Johnson-Laird, 2012; Johnson-Laird, 1983). Evidence for this view of negation processing has been obtained from studies showing that perceptual aspects of the embedded proposition are primed when negations are presented: for instance, picture-naming studies using a priming paradigm suggest that comprehenders must at some stage switch their attention from simulating the negated situation to simulating the actual described situation (Kaup & Zwaan, 2003; Kaup, Lüdtke, & Zwaan, 2006; Kaup & Lüdtke, 2007); further discussion and evidence for this perspective is provided by Hasson and Glucksberg (2006), who used evidence from negated metaphors to avoid issues with lexical priming in concluding that comprehenders switch to representing a full negative proposition

around 500 to 1,000 ms after it is read, and also by Anderson, Huette, Matlock, and Spivey (2010) and Tian, Breheny, and Ferguson (2010). However, a simulation model of negation processing does not necessarily mean that a two-stage account is necessary. Huette and Anderson (2012) present a recurrent network model proposing that a simulation account can handle at least some types of negation in a single step, without the need to simulate the world described by the embedded proposition or to use any kind of logical tag, as in schema-plus-tag accounts.

It became clear relatively early that any processing slowdown, two-step process, or other type of difficulty caused by negation is not constant across all contexts or situations in which negations can be presented. Wason (1965) showed that participants were quicker to verify negative descriptions of a set of items when they were first described in terms of an “exceptional item and a residual class”, rather than in terms of a larger set and a smaller set. Early links made between the philosophical and semantic approaches described in section 1.2.1 above and early psycholinguistic findings led to the suggestion that some of the effects observed might be accounted for by the specifics of the purposes for which negation is used naturalistically (Apostel, 1972a; Wason, 1972). These observations can be summarised by Apostel’s suggestion that the use of a negation *not-q* tends to mean or imply “incompatible with *q* but in the neighbourhood of *q*” (Apostel, 1972b). Along similar lines, De Mey (1972) makes the case that negation can only be made sense of as an invitation for the comprehender to shift their attention, and thus to understand it fully in psycholinguistics, it is crucial to present fully fleshed-out contexts: “‘Natural’ negation only involves objects or elements a speaker or a listener is attending to. Negation then appears as a ‘meta-operator’, instructing the listener to attend no longer to a possibility he is considering. It makes no sense to instruct a listener to suppress a thought he is not considering or an idea he is not having” (p. 149).

Thus, a case can be made that a critical factor in determining the approach a comprehender must take to processing negation is the extent to which a clear concept for the communicated state of affairs is both defined and available. This defining can take place on one of several levels: for example, in the lexical semantics of the words chosen to express the negation, or on a pragmatic level where the context of an utterance constrains the meaning of a lexically ambiguous word to make the current negation of it obvious. The first case is what we would ordinarily describe as bipolar negations: even in the absence of any particular context, there is a clear opposite (e.g., *clean*, the negation of which clearly implies *dirty*). The second

case involves negated forms that may be unipolar if presented in isolation, but that present a clear opposite in the specific context given. For instance, *not creative* could imply *dull* in a context like “John never came up with new ideas. He was simply not creative” or *unartistic* in a context like “John never made much progress in his art class. He was simply not creative”. The pragmatics of the context clearly select from among these possibilities. This may have the effect, in processing terms, of transforming a unipolar negation into a bipolar negation. Löbner (2000) discusses pragmatic and cognitive aspects of this type of polarisation, arguing that this process can be understood as a simplification strategy for conveying useful insights about the real world without introducing too much complexity. Thus, the context of a negative statement can set up a “contrast frame” or “contrast set” that aids the comprehender in interpreting what is conveyed by it, creating a locally constructed polarity contrast defined by the relevant discourse presuppositions.

This type of argument leads towards the hypothesis that an adequately realistic usage of negation, in a pragmatically enriched discourse context, may present a meaningfully more straightforward processing problem — to a greater extent than is true for affirmatives. Approaches such as those by Glenberg, Robertson, Jansen, and Johnson-Glenberg (1999) and Giora (2006) present the modern view that there is arguably no fundamental processing asymmetry between affirmatives and negations when the higher-level pragmatic functionality is taken into account; similarly, Lea and Mulligan (2002) show that where negative information is useful in making deductions, it is readily incorporated.

These views are supported empirically by more recent studies finding that, given adequate pragmatic context and support for the use of negation, incrementally-formulated predictions made by comprehenders can take negation into account. These include Nieuwland and Kuperberg’s (2008) finding that the pattern of ERP findings observed by Fischler et al. (1983) can be reversed when pragmatically felicitous negative sentences are presented (e.g., “With proper equipment, scuba diving is not very *dangerous* and often good fun”), as well as similar work by Dale and Duran (2011) using a computer mouse-tracking methodology and Orenes, Moxey, Scheepers, and Santamaría (2016) using an eye-tracking approach. These findings are discussed in more detail in the introductory sections of the relevant chapters below.

Taking together these empirical findings that pragmatic context is a critical factor in what comprehenders can do with an incomplete proposition involving negation and the other work, described above, on the processing mechanisms

involved (e.g., creation of a relevant contrast set that allows the negation and negated concept to be treated as a unit with coherent meaning), an explanation can be proposed for why pragmatic felicity has such a major impact on the processing of negation. Increasing the felicity of a negative statement, through the presentation of a particular discourse context, renders upcoming content more predictable, which has the effect of making negations more easily “solved” under this type of processing. This means that, once a negating element such as *not* is encountered, it is both easier to interpret this element and easier to predict what is likely to follow it (i.e., to be negated), since the possible space of likely upcoming communicative messages has been constrained.

Under this view, prediction forms a critical component of why increasing the pragmatic felicity of a piece of linguistic input has such effects on the processing of negation. The hypothesis that pragmatic and semantic predictability constitutes the underlying factor governing the extent to which negation can be processed incrementally is the overriding driver of the work presented here, which is briefly introduced in section 1.3 below.

1.3 Introduction to experimental work

The following chapters present a set of experimental approaches to investigating the specific role of predictability — as divorced from pragmatic felicity — in the incremental processing of negation. To achieve this, the general approach is to manipulate episodic contextual information supplied along with each instance of linguistic input. For example, in Chapter 2, brief animations are presented along with sentences describing the events that occur (or do not occur) in them, such as “The wizard didn’t raise the position of the basket”. This strategy allows for the presentation of only sentences that are pragmatically licensed by the context, without relying on real-world semantic knowledge and associations to support the licensing; furthermore, by varying the context presented, the same sentence can be employed in conditions with different levels of predictability. This type of design is explained in more detail in the relevant sections of each experimental chapter.

The overarching hypothesis is that, even in the case of sentences that are all equally pragmatically felicitous, predictability should be expected to have an impact on the extent to which negative sentences can be processed as readily as affirmatives. Specifically, when predictability is high, negations should be expected

to behave very much like affirmatives, in that measures tapping into comprehenders' predictions about upcoming material should be expected to show that they take into account all available information, including the presence of negation; conversely, when predictability is low, negations should be expected to behave (at least during early stages of processing) less like affirmatives, with predictions formulated less readily and perhaps actively containing mistakes, demonstrating that they have been formulated without taking negation into account.

Several methodological approaches are employed to test this broad hypothesis by accessing such mid-processing predictions. Chapter 2 presents an EEG experiment measuring modulation of the N400 component. Next, three computer mouse-tracking experiments are presented, in which participants' mouse trajectories while making truth value judgements (Experiment 1, Chapter 3) and selecting picture-based sentence completions (Experiments 2 and 3, Chapter 4) are examined. Finally, a preliminary eye-tracking experiment (Chapter 5) explores the scope for use of this type of paradigm in investigating the impact of various factors, including predictability, on the incremental processing of negation. Chapter 6 provides a general discussion and conclusion drawing together the evidence provided by each of these experimental approaches, in light of the existing literature.

Chapter 2

EEG Study

2.1 Introduction

2.1.1 Electroencephalography in language research

Electroencephalography (EEG) is a technique in which electrical signals generated by the activity of the brain's neurons are measured, generally using electrodes placed on the scalp. Unlike imaging techniques such as fMRI, EEG provides only impoverished information on the location of brain activity, but one of its advantages is extremely high resolution in time (e.g. Luck, 2005). By correlating the presentation of stimuli with variations in the characteristics of the signal, researchers can observe the effects of manipulating these stimuli. Over time, various useful components of this time-locked response (the event-related potential, or ERP) have been identified; these have been found to represent responses to particular aspects of certain stimuli, and thus, their magnitudes can usefully be employed as dependent measures. Some of these components, which may be both event-preceding and event-following, are reviewed by Coles and Rugg (1995).

Several ERP components have proven particularly relevant to language-related research questions. One of these is the N400, which was first reported by Kutas and Hillyard (1980) to be elicited by words that were semantically incongruous in a sentence context, e.g., "He spread the warm bread with *socks*". The N400 is a negative-going component that occurs late after the onset of stimulus presentation, usually with a peak latency around 400 ms. The component has been the subject of extensive debate (reviewed by Kutas & Federmeier, 2000, 2011) as to its properties, the nature of the underlying cognitive processes that it can be taken to reflect, and how variations in its amplitude should be interpreted. The component is

modulated primarily by semantic violations of various kinds: according to Kutas and Hillyard (1984), its magnitude is roughly the inverse of a critical word's cloze probability (the relative frequency with which readers select the word in question as a candidate for the next word when presented with the partial sentence up to that point). The component has also been found to be sensitive to the distinction between within-category and between-category violations. For example, in a context ("They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of...") in which *palms* is expected, *tulips* (between-category violation) elicits a larger N400 than *pinos* (within-category violation), even though both completions are an equally poor fit (Federmeier & Kutas, 1999).

It should be noted that Rayner and Clifton (2009) offer some cautions for the use of ERPs in language research, given the speed and incrementality of linguistic processing. To an extent, they point out, the focus on later components (which may completely miss some of the relevant processes) may arise from the use of rapid serial visual presentation (RSVP), in which a sentence is presented visually word-by-word, to avoid the timing and artefactual (eye movement) problems that may arise in such experiments from the use of presentation methods like spoken presentation or natural reading. They note that the use of spoken presentation may produce results that differ in important ways, particularly at earlier stages of processing.

2.1.2 Using the N400 to investigate incrementality and negation

As outlined in the General Introduction, the extent to which language comprehension proceeds incrementally, with continuously updated, active predictions, has been a topic of extensive research. ERPs provide a highly useful measure for this line of investigation, and the N400 component is especially relevant because of the way it can be used as a proxy for the extent to which part of a sentence was semantically expected or predicted by the comprehender, or the ease with which it is integrated into the comprehender's representation of the sentence. Thus, it can be used to answer questions about various levels of linguistic processing. For example, van Berkum, Zwitserlood, Hagoort, and Brown (2003) presented sentences containing critical words with opposite meanings, in which either was equally acceptable given only the local content of the sentence (e.g., "Jane told her brother that he was exceptionally quick/slow"). They found that when additional discourse content was introduced preceding this sentence, rendering one of the alternatives contextually appropriate and the other contextually anomalous, an N400 effect

appeared in which a larger N400 was evoked by the anomalous word, showing that the wider discourse is integrated rapidly into the comprehender's interpretation of and predictions for the sentence. Hagoort and van Berkum (2007) review a number of effects relating to the integration of more general world knowledge and speaker knowledge; the overall picture suggests that contextual knowledge and the meaning of the sentence are immediately and continuously used to update the comprehender's interpretation of an utterance, in contrast with two-step models suggestion that the meaning of each part of a sentence must be computed before integrating it into the surrounding structure. More recently, in the domain of syntactic interpretation, Chow et al. (2016) used stimuli in which the syntactic roles of arguments were exchanged to show that comprehenders' predictions about upcoming material (as indexed by their N400 responses) are influenced systematically by the syntactic roles of words, and not merely by semantic associations with a "bag of words" regardless of their relationships to each other.

Tackling a topic with more direct similarities to the subject of particular interest here, Urbach and colleagues used N400 amplitudes to investigate whether quantifiers are interpreted incrementally, a question that has many commonalities with questions about the processing of negation (such as the relevance of scope and the possibility for various types of two-stage processing). Urbach and Kutas (2010), using sentences of the form "Most/few farmers grow crops/worms as their primary source of income", found that the magnitude of the N400 component was reduced in response to *worms* (and increased for *crops*) when *most* was replaced with *few*, but a full reversal of the effects was not observed, suggesting that incorporation of the full meaning of quantifiers into the sentence representation is not immediate, but is also not fully delayed. A later study including pragmatically-supportive discourse contexts for similar sentences (e.g., "Alex was an unusual toddler. Most/Few kids prefer sweets/vegetables...") suggested that the presence of such contexts induced a full reversal of the effects depending on which quantifier was presented, once again highlighting the relevance of the pragmatic context to incremental interpretation (Urbach, DeLong, & Kutas, 2015). In this case, a quantifier with negative force was sufficient to reverse participants' expectations about upcoming material.

Turning to N400 research specifically examining the processing of negation, an early study by Fischler et al. (1983) investigated the effects on N400 amplitude of presenting true and false negative sentences such as "A robin is (not) a bird/tree". They found that sentences of this type elicited a large N400 to *tree* and a reduced

N400 to *bird* regardless of whether *not* was present in the sentence (and therefore, whether it was true or false), concluding that the N400 therefore reflects the strength of the semantic associations between the critical word and preceding information (in this case, *bird* is closely associated with *robin*, whereas *tree* is not), rather than the truth value of the sentence, and that the embedded (negated) proposition in such sentences may be interpreted before applying the negation.

More recent work, however, has produced more mixed findings, identifying various scenarios in which the N400 in negative contexts can in fact be modulated by truth value or offline cloze probability, rather than primarily by semantic associations. For example, Lüdtke, Friedrich, de Filippis, and Kaup (2008) presented German sentences describing a positional relationship between two objects, e.g., “Vor dem Turm ist ein / kein Geist” (In front of the tower there is a / no ghost). Each sentence was followed by a picture that was either congruent or incongruent with the sentence: in the example case, there was either a ghost or a lion in front of a tower. They found that when the picture was presented after a short delay (250 ms) following the sentence, the amplitude of the N400 (as well as behavioural responses) elicited by the picture was simply a function of whether the object present in the picture was the one mentioned in the sentence or not, regardless of whether the picture was congruent with the actual meaning of the sentence. However, when the delay was much longer (1,500 ms), main effects of truth value and negation were also observed (i.e., the extent to which the N400 response was governed solely by whether the figure present in the picture matched the entity mentioned in the sentence regardless of negation was reduced), suggesting that the effects of the negation on participants’ predictions about the image could be added to the priming effects, partially cancelling them out, given sufficient time.

Ferguson, Sanford, and Leuthold (2008) measured N400 responses as well as eye-movements in response to semantically congruent or anomalous critical words in sentences preceded by counterfactual scenarios in which real-world expectations were negated but no explicit alternative scenario was constructed: for example, “If cats were not carnivores they would be cheaper for owners to look after. Families could feed their cat a bowl of carrots / fish and it would gobble it down happily”. Based on both measures, they drew similar conclusions to Lüdtke et al. (2008), finding that the introduction of a negated-world (as opposed to a real-world) context was insufficient to immediately update comprehenders’ predictions about the input, which was instead tested against real-world knowledge. This finding does not constitute strong evidence that negation was the sole cause of

this incomplete updating, as counterfactual sentences couched in an affirmative manner were not tested, but it was taken by the authors to show that negation does not always act purely to suppress concepts related to the negated entity. They propose that instead it acts as a prompt to initiate a search for specific alternatives to the negated concept.

Nieuwland and Kuperberg (2008) noted that one potential confounding factor in the original sentences used by Fischler et al. (1983) was that the negative sentences tended to be pragmatically infelicitous: that is, there was no particular discourse or other context making the production of sentences like “A robin is not a tree” relevant. Since negations are normally used to deny or otherwise contradict a proposition that may already be under consideration for some reason, they are more sensitive than affirmatives to licensing conditions rendering them pragmatically felicitous or infelicitous. Nieuwland and Kuperberg tested the hypothesis that the infelicity of these sentences formed part of the explanation for the the failure of their truth value to modulate the N400 by crossing pragmatic felicity with truth value. They presented affirmative or negative and true or false sentences that were rated as either natural (e.g. “With proper equipment, scuba diving is/isn’t very *safe/dangerous* and often good fun”) or unnatural (e.g. “Bulletproof vests are/aren’t very *safe/dangerous* and used worldwide for security”). For unnatural (i.e., pragmatically unlicensed) sentences, they replicated Fischler et al.’s original (1983) findings, but for licensed sentences, they found that N400 responses followed the same pattern for negatives as for affirmatives: that is, a critical word rendering the sentence false elicited a large N400, and one rendering the sentence true elicited a reduced N400, regardless of whether negation was involved or not. They concluded that, provided that negative sentences are used to present information that is pragmatically licensed by a real-world context, the presence of negation poses no particular obstacle to the incremental updating of a sentence representation that incorporates all available information.

More recently still, Xiang, Grove, and Giannakidou (2016) examined N400 responses to *ever*, a negative polarity item (NPI) that must be licensed by a negative context. The authors compared NPIs licensed by asserted negation (e.g. “The teacher brought a tarantula to class. *No/Few/Only* third-graders had ever seen one before”) and by implied negation (e.g. “She was *surprised* that third-graders had ever seen one before”). They found that, in all of these cases, the NPI elicited a smaller N400 than in the case of use of an NPI without a licensing context (e.g. “*Third-graders had ever seen one before”), further supporting the view that the

presence of a (pragmatically appropriate) negation can incrementally influence comprehenders' preparation for or responses to upcoming material. In this case, the predicted or readily interpreted material was dependent on a syntactic relationship (licensing) arising from the availability of semantic or pragmatic information, illustrating the multiple levels over which negative information can be accessed and incorporated.

2.1.3 The role of pragmatic licensing and predictability

Nieuwland and Kuperberg's (2008) study is one of several, across various methodologies, finding that pragmatic licensing is an essential predictor of whether comprehenders can incrementally incorporate negation into their interpretation of a sentence and update their predictions accordingly. However, it is not clear exactly why pragmatic licensing has this effect, since people are easily able to judge the truth value of negated sentences whether they are licensed or not. One possibility is that pragmatic felicity is intrinsically related to predictability. This relates to the abovementioned "search-for-alternatives" described by Ferguson et al. (2008), in that the scope of this search can be narrowed considerably by contextual information that increases the predictability of the sentence. In the absence of pragmatic context providing information on *why* a particular proposition specifically has been selected to be negated (e.g., a correction of an interlocutor's mistaken belief), it is very difficult for a comprehender to make predictions about upcoming material in a sentence that incorporate the interpretation of the negation. In other words, there is no completion with a particularly high cloze probability. The case can be made that the probability distribution over possible subsequent words tends to look rather different in the case of affirmative sentences: even in the absence of a similar type of context to that which renders negations more predictable, there are more concepts that are associated with the information available as part of the utterance so far than there are concepts that are *not* associated. Thus, although there may be no single completion with an extremely high cloze probability, the overall distribution is much more likely to be centred on a relatively small number of relatively high-probability possibilities.

In fact, in many cases, among the best predictions a comprehender can make during incremental interpretation of a negative sentence might be those that would seem to render the sentence false, because in the absence of any other context, the speaker might be intending to deny a commonly accepted belief that they claim is a misconception ("The food that Marie Antoinette is supposed to have suggested

the peasants should eat wasn't cake, actually — it was brioche.”). That is, in the absence of any richer context, *bird* is not such a poor bet to complete the sentence “A robin is not a ...”, even though it produces a falsehood; and it is much easier to predict than *tree*, as without any further information for the context in which the statement is being made, there are virtually infinite options for final words that could complete the sentence to make it equally true. Including a richer pragmatic context has the effect of narrowing the field of possibilities for the pragmatic force that the speaker intends their negative utterance to have, perhaps identifying a relevant reference class, thus allowing the possibility of making predictions that take into account the negation.

The research question tackled in the present study was whether this predictability can in fact be regarded as the “active ingredient” in the pragmatic felicity: that is, whether predictability influences the extent to which negation can be processed incrementally, in the absence of any associated variation in pragmatic felicity.

2.1.4 The present study

In view of the fact that few previous experiments have examined predictability separately from pragmatic factors as a possible factor in the ease and incrementality of processing of negations, an experimental paradigm that would permit this was designed. In particular, the aim was to present participants with only pragmatically licensed sentences, while varying the predictability of a critical word.

To achieve this, episodic scenarios were presented visually using brief animations; these constructed temporary relationships between objects and predicates (in this case, actions applied to the objects by a character) in a highly constrained, simulated world. Each animation was accompanied by a sentence describing what had happened in the animation, mentioning only actions and objects that had featured in it. These sentences could be true or false. Although information that is only relevant episodically may not always be added in a lasting way to real-world long-term knowledge, it has been found that the N400 can be modulated on the basis of this type of information. In addition to the studies described above that made use of episodic information, this has been tested specifically, for example by Fischler, Childers, Achariyapaopan, and Perry (1985), who presented participants with a list of facts about the supposed occupations of fictional characters (e.g. “Mary is a lawyer”) prior to EEG recordings of their responses to true and false sentences based on this information. A larger N400 was elicited by false (i.e., not in accordance with the provided information) statements of the form than by

true statements, even though there was no pre-existing or real-world association between “Mary” and “lawyer” and the participants presumably did not update their real-world judgements about the occupations of any people. Contextual information can also override responses that would be strongly evoked in the absence of such information, for example in the case of an animacy violation. This was demonstrated by Nieuwland and van Berkum (2006), who found that through the presentation of sufficient preceding context (a story about a peanut singing a song about his girlfriend), a larger N400 component could be elicited in response to the normally acceptable and somewhat predictable sentence “The peanut was *salted*” in comparison to the normally unacceptable and surprising sentence “The peanut was *in love*”. On the basis of such findings, it was anticipated that, in the present experiment, true and false sentences would modulate the N400 in the same way as sentences whose truth value is based on long-term, real-world semantic memory.

This approach has several advantages over presenting sentences that are true or false based on general, real-world knowledge. First, it is not always possible to be certain that assumed real-world information is shared by all participants, or to the same degree. Even if a statement is universally agreed to be true, its truth might be more obvious or more readily accessible to some participants than to others, for example if they are more or less familiar with the domain in question. In contrast, presenting an episodic context clearly rendering the statement true or false immediately prior to the statement ensures that the truth value of the latter is equally unambiguous and equally accessible to all participants. Second, it is difficult to control for the effects of the actual linguistic input (e.g., the lengths of words or their associations with other words) when presenting statements based on real-world knowledge, as the sentence must be formulated around the real-world information, and true and false sentences must obviously differ from one another, such as through the use of contrasting predicates (e.g., in Nieuwland & Kuperberg, 2008: “With proper equipment, scuba diving is very *safe* / *dangerous* and often good fun”). When the truth or falsity of a sentence is instead based on episodic information, it is possible to present exactly the same sentence to different participants and (through the use of different episodic contexts) have it represent a true sentence in one case and a false sentence in the other. The same is true for the other independent variable manipulated in the present experiment, namely predictability. Thus, this design allowed the presentation of linguistic stimuli that were perfectly controlled for content across participants, while their

truth value and the predictability of the critical word were manipulated by altering the accompanying animation.

Third, and most crucial to the aim of the present experiment, this approach enabled the manipulation of predictability independently of pragmatic felicity or licensing. When real-world information is used to vary the predictability of a critical word in a sentence, its pragmatic felicity tends to vary in the same direction (and vice versa), because, as discussed above, highly felicitous sentences are likely to be very predictable. For example, although Nieuwland and Kuperberg (2008) did not obtain predictability norms in Nieuwland and Kuperberg (2008), intuition suggests that the critical word was considerably more predictable in their pragmatically licensed sentences (“With proper equipment, scuba-diving is very...”) than in the unlicensed sentences (“Bulletproof vests are very...”). (In both these cases, the completion was *safe*.) Although the correlation between felicity and predictability in such sentences is not perfect, it is rather difficult to assess them independently in this type of paradigm.

In a related experiment with quantifiers, Nieuwland (2016) did collect measures of predictability in the form of cloze probabilities: for example, in the stimulus set “Many/few gardeners plant their flowers during the *spring/winter* for best results”, the critical word (in this example, the name of the season) was highly predictable in the case of the true pair of sentences. The authors found that sentences with high cloze probability showed a pattern similar to that associated with sentences in the pragmatically licensed condition in Nieuwland and Kuperberg (2008), with N400 amplitude at the critical word modulated in the same way by truth value regardless of whether a positive or negative quantifier was used, suggesting that participants were able to incrementally incorporate information from quantifiers into their interpretations of such sentences, whereas low cloze probability sentences elicited an interaction between truth value and quantifier type. This constitutes further evidence that predictability forms an important component of such incremental processing effects, although because this paradigm also relied on world knowledge to manipulate truth value, predictability could not be completely differentiated from variables such as felicity and lexical co-occurrence of the words in each sentence.

In contrast, manipulating the truth value of sentences using episodic information means that predictability is not only independent of felicity, it is also tightly and quantifiably controlled. Furthermore, because all the sentences used in the present experiment were rather predictable (the critical word could be predicted

with a probability of 1 in predictable conditions and 0.5 in unpredictable conditions), it was expected that they would elicit particularly large N400 effects, because comprehenders would be able to make strong predictions that would be either clearly met or clearly violated. In turn, this anticipated large effect size could make the effects of polarity easier to detect, by providing scope for an interaction that could also have a larger effect size.

2.1.5 Design and hypotheses

The independent variables manipulated on each trial were polarity of the sentence and predictability of the critical word. Truth value of the sentence was also manipulated in order to compare the N400 response to true and false sentences in each condition. Thus, the experiment employed a 2 (polarity: affirmative or negative) \times 2 (truth value: true or false) \times 2 (predictability: high or low) design. All variables were manipulated within participants, with counterbalancing for extraneous variables, such as features of the stimuli (animations and sentences) carried out between participants (see Methods for full design details).

In the case of highly predictable sentences, a replication of Nieuwland and Kuperberg's (2008) findings was anticipated: namely, that processing would be equally easy and incremental for both affirmatives and negations (all being pragmatically licensed). Reflecting this, large N400s to false sentences and reduced N400s to true sentences were expected, regardless of polarity. However, in the case of (relatively) low predictability, it was expected that even these pragmatically licensed sentences would present processing difficulties in the presence of negation, evoking similar effects to those seen by Fischler et al. (1983) in the case of pragmatically unlicensed sentences. Reflecting this, an attenuation or even reversal of the pattern predicted for highly predictable sentences was predicted in the case of negations. Specifically, in the case of low predictability sentences, large N400s to false sentences and reduced N400s to true sentences were expected for affirmatives (as with high predictability sentences), but for negations, in the case of the most extreme form of this effect, reduced N400s to true sentences and large N400s to false sentences could be expected.

Thus, the hypotheses were as follows: (1) a main effect of truth value would be observed, with true sentences overall eliciting reduced N400s; (2) a truth value \times polarity interaction would occur, in which the N400 reduction for true sentences value would be attenuated or even reversed in the case of negations; and (3) a truth value \times polarity \times predictability interaction would occur, in which the

aforementioned reversal of the truth value effect for negations would be found to operate more strongly in the case of low predictability.

2.2 Methods

2.2.1 Materials

Stimuli

Stimuli consisted of short animated scenes, each paired with a sentence describing an aspect of the corresponding scene. In each animation, three everyday objects (from a set selected for their high recognisability and nameability) appeared and a “magical” character (e.g., wizard or fairy) interacted with them by casting two different spells (Figure 2.1). For each spell, the intended outcome was represented in the form of a thought bubble showing that the character was thinking about a particular action. Following the thought bubble, the spell was directed towards all three objects, as shown by star-shaped “sparks”, accompanied by a sound effect. The outcome of the spell was displayed for each object in turn, from left to right. First, the image of the object blinked three times to indicate that its outcome would now be displayed. In the case of each object, the spell could either succeed or fail. If it succeeded, a chime sound effect was played and the animation showed the action being carried out for that object (e.g., a square dropped into the scene over it, or the colour was gradually drained from the object). If it failed, a buzzer sound effect was played and nothing changed about the object. This process was displayed for each object and for each spell. After the second spell, the final frame from the animation (with all actions completed) remained on the screen as a still for 500 ms. In total, each animation lasted approximately 15 s.

For animations associated with critical trials, one of the two spells always succeeded for exactly one of the objects (the critical object) and failed for the other two, while the other spell always succeeded for the two non-critical objects and failed for the critical object. For animations associated with filler trials, these constraints did not apply, and the outcome of each spell (success or failure) was decided independently for each object. Thus, on filler trials, objects could be affected by both actions, or by neither action.

For critical trials, each animation could be paired with eight different sentences, of the form “The fairy deposited / didn’t deposit a square around the leaf in that scene” (Table 2.1). The sentence could be affirmative or negative (deposited vs.

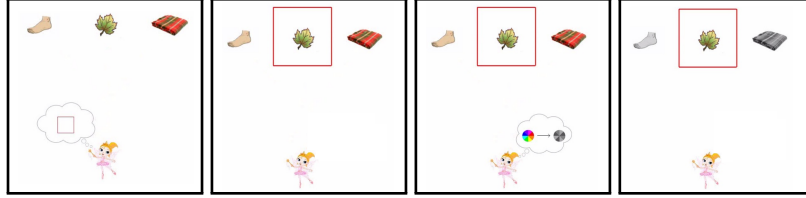


Figure 2.1: EEG experiment visual context stimulus: example frames from an animation presented before a critical sentence. In the first panel, the character thinks about performing the “deposit a square around” action. This action is attempted on each object in turn; in the second panel, it can be seen that it has succeeded for the leaf, but failed for the foot and the blanket. In the third panel, the character thinks about performing the “fade the colour of” action. Again, this action is attempted on each object in turn; in the fourth panel, it can be seen that it has succeeded for the foot and the blanket, but failed for the leaf. This state of affairs sets the stage for each of eight possible critical sentences (see Table 2.1).

	Affirmative	Negative
High predictability	<i>True:</i> The fairy deposited a square around the leaf in that scene.	<i>True:</i> The fairy didn’t fade the colour of the leaf in that scene.
	<i>False:</i> The fairy deposited a square around the foot in that scene.	<i>False:</i> The fairy didn’t fade the colour of the foot in that scene.
Low predictability	<i>True:</i> The fairy faded the colour of the foot in that scene.	<i>True:</i> The fairy didn’t deposit a square around the foot in that scene.
	<i>False:</i> The fairy faded the colour of the leaf in that scene.	<i>False:</i> The fairy didn’t deposit a square around the leaf in that scene.

Table 2.1: EEG experiment linguistic stimuli: all possible sentences that could follow the example animation shown in Figure 2.1.

didn't deposit), true or false (in reference to the animation), and predictable or unpredictable. In the case of predictable sentences, at the point immediately preceding the object name, there was only one possible object (the critical one in the animation) whose name could complete the sentence to make it true. In the case of unpredictable sentences, there were two possibilities (either of the non-critical objects). All sentences ended with post-critical material of at least two words, to avoid end-of-sentence effects in the ERP to the object name. For filler trials, the sentences took the same form, but because there was no critical object in the corresponding animation, they did not generally fall into the categories of "predictable" or "unpredictable".

All sentences were considered to be pragmatically licensed, because each sentence referred to an action and an object that had featured in the episodic context (nothing was mentioned "out of the blue"); furthermore, even when an action had not applied to a particular object, an attempt had always been made by the character to apply it, meaning that it was felicitous to deny that this had happened. Nevertheless, the predictability of the critical word could still be manipulated. This was achieved by the fact that, on critical trials, one of the actions applied to two different objects, while the other action applied only to a single object, meaning that in some conditions, only a single object name could complete the sentence to make it true, whereas in others, either of two object names would accomplish this.

In total, 176 animations were created. Of these, 56 were associated with filler trials, which were completed by every participant. For the remaining 120 (associated with critical trials), two different versions of the animation were created (for a total of 240), in which the critical object exchanged roles with one of the non-critical objects. This ensured that pairs of trials with identical linguistic stimuli could be presented (to different participants) in each predictability condition, by using the same sentence accompanied by each version of the animation, thus controlling for any item-related effects of the identity of the named object (for example, its salience or nameability).

Each of these 120 pairs of animations was associated with eight possible sentences, producing 16 possible trials, for a total of 1,920 possible trials. The allocation of trials to participants was counterbalanced such that no participant saw more than one trial from a set of 16, while each participant saw an equal number of trials in each condition and each possible trial was presented exactly twice (to different participants).

Animations were created, stimuli displayed, and participants' responses recorded using the Presentation software (Neurobehavioral Systems), which was also used to output timing markers to the software recording the EEG signal (see below).

Memory test

Sets of sentences were also constructed for use in a memory test following the main experiment. The stimuli for each participant consisted of 48 sentences of the same form as the sentences presented in the main experiment, each being associated in the same way as in the main experiment with one of the animations that the participant had seen. Half of the sentences had in fact featured in the main experiment in the form used in the memory test, whereas the other half were new sentences that the participant had not already seen. Additionally, half the sentences were true (with reference to the relevant animation seen by the participant) and the other half were false; and half were affirmative and half negative. All memory test sentences would have fallen into the low predictability condition in the main experiment.

EEG recording apparatus

Scalp EEG was recorded continuously from each participant during the main experiment, using an Acticap elasticated cap (Brain Products) with 32 active Ag/AgCl electrodes referenced online to FCz (Figure 2.2). Data from each electrode were sampled at a rate of 1,000 Hz, amplified using a BrainAmp DC amplifier (Brain Products), and recorded using the Vision Recorder software (Brain Products) on a separate computer to the one used to present stimuli and record responses. The recordings were also marked, using port transmissions from the stimulus-presenting computer, to indicate the timings of critical events, including the onset of each word presented in the sentence and the participant's behavioural responses to the task.

During EEG recording, the participant was seated in a sealed Faraday cage to reduce electrical noise in the EEG signal.

2.2.2 Participants

A total of 32 participants, 25 female, aged 18–23 years ($M = 19.6$, $SD = 1.3$) were recruited from in and around the University of Bristol community. All were native speakers of English and most were monolingual; three spoke a second language

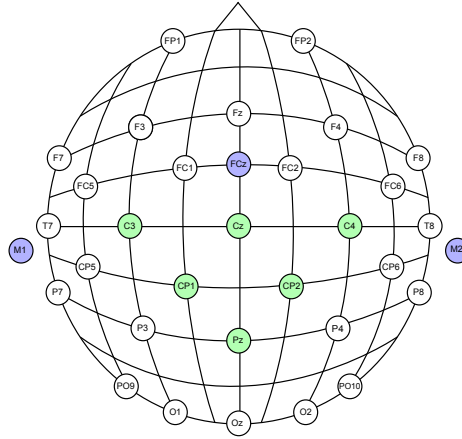


Figure 2.2: Electrode layout during recording in EEG experiment. M1 and M2 were placed on the left and right mastoids in place of electrodes at TP9 and TP10. Fz was used as a ground and all channels were referenced online to FCz. For analysis, data from each channel was rereferenced to the mean of M1 and M2. Green electrodes represent the centroparietal region of interest.

(Mandarin, French, or Hebrew) to an advanced or fluent standard. Three were left-handed and the handedness of one participant was not recorded.

All participants gave their informed consent to participate and were compensated for their time with course credit or a small payment. Ethical approval for the experiment was granted by the University of Bristol Faculty of Science Ethics Committee (ethical approval code 18541).

2.2.3 Procedure

Prior to providing their informed consent to participate, each participant was familiarised with the EEG laboratory and materials, including the Faraday cage chamber, cap and electrodes, syringes and needles used for gelling, and intercom through which they could communicate with the experimenter. The correct cap size for the participant was selected based on head measurements and 32 electrodes fitted to the cap according to the standard layout, with the exceptions of the electrodes designated TP9 and TP10 (Figure 2.2).

Following this process, the participant was seated in a chair a comfortable distance from the screen inside the Faraday cage. The appropriate cap was fitted carefully, with measurements taken to ensure that it was positioned correctly, and secured using a chin-strap. Electrodes TP9 and TP10 were placed underneath the

cap at the left and right mastoids (M1 and M2), respectively, and affixed to the skin using microporous tape. Saline gel was then applied to the hair and scalp at the site of each electrode to ensure good conductivity between the scalp and electrode. In most cases, impedances at each electrode were 5 k Ω or below. Impedances could not be recorded for three participants due to a software error, although visual inspection of the signal suggested that they were at an adequate level.

Once gelling was complete, the participant was instructed on the task and asked to avoid blinking and moving to the greatest extent possible during presentation of sentences. They wore headphones, placed over the electrode cap, through which sound effects forming part of the animations were presented, and held a gamepad controller that was used to give responses. The door to the Faraday cage was sealed and the experiment began when the participant was ready.

On each trial, the participant watched a brief animation (see Materials section above). Following the animation, the question “True or false?” was displayed in the centre of the screen for 300 ms, followed by a blank screen for 200 ms. Each word in a sentence was then displayed individually in the centre of the screen for 300 ms, with a 200 ms blank screen between each word. After the final word, the prompts “True” and “False” appeared on the left and right sides of the screen, respectively, with arrows indicating that the participant should respond using the corresponding button on the gamepad. No feedback on responses was provided. Following the response, the screen displayed a message inviting the participant to begin the next trial by pressing a button on the game pad.

Trials were presented in eight blocks, each consisting of 22 trials. Participants were invited to take a longer break to move around (while remaining seated and wearing the cap), rest, etc. between blocks. In total, each participant completed 176 trials, presented in a random order, of which 120 were critical (15 in each experimental condition) and 56 were fillers.

After the main experiment, the cap was removed and the participant was given the opportunity to walk around and take a longer break. Subsequently, they were asked to complete a memory task in which they were presented with sentences that either had or had not been included in the main experiment. Sentences were presented visually in the centre of the screen using the MouseTracker software (Freeman & Ambady, 2010), and participants clicked “Read” or “Not read” response buttons to indicate for each sentence whether they believed they had encountered it during the main experiment. Responses and mouse trajectories were recorded, but only response accuracies were explored.

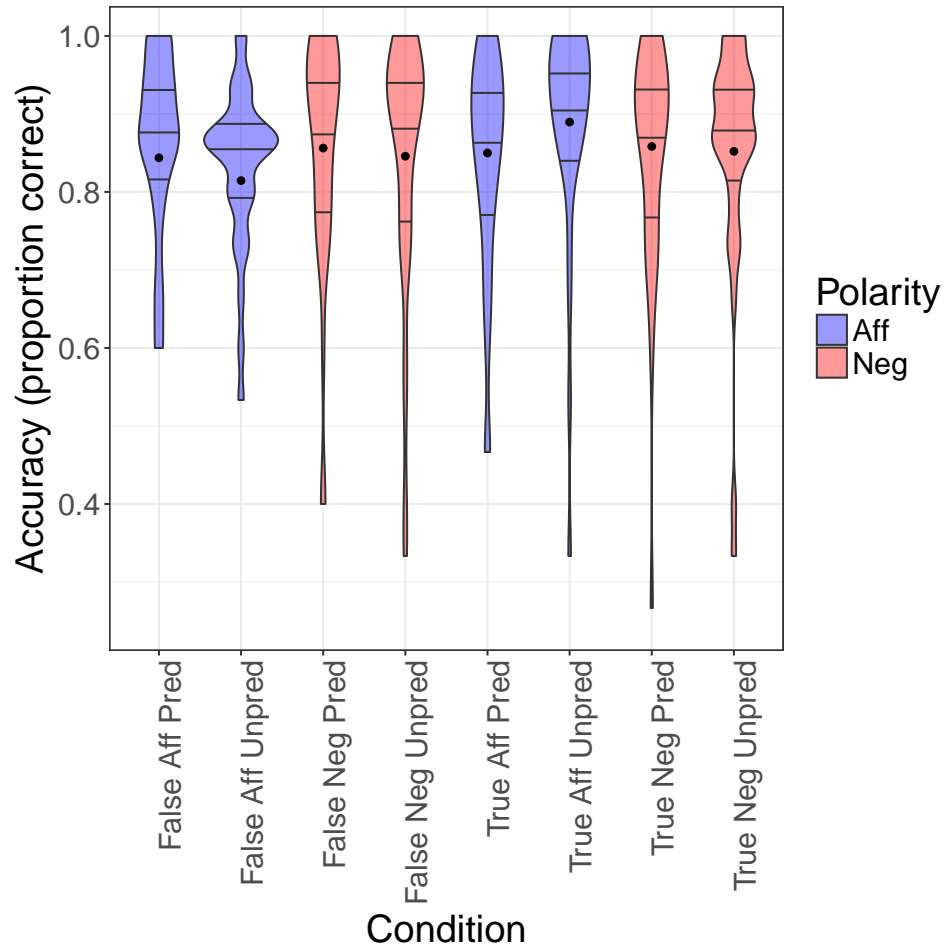


Figure 2.3: EEG experiment response accuracy violin plot: proportion of correct responses in the truth value judgement task for each condition. Dots represent the mean; horizontal lines represent quartiles.

2.3 Results

2.3.1 Behavioural response accuracy

Participants' mean response accuracy across all conditions in the truth value judgement task was 88%. Figure 2.3 illustrates accuracy for each condition. Differences between conditions were analysed as follows.

To test for effects of truth value, polarity and predictability on response accuracy, mixed effects logistic regression models (with a binomial family error distribution and logit link function) were constructed over proportion of correct responses using the glmer function of the R package lme4 (Bates, Mächler, Bolker, &

Walker, 2015). The full model included terms representing fixed effects of all three independent variables, their two-way interactions, and a three-way interaction among all of them, as well as a random effect by participant of truth value. (Models with a more complex random effects structure failed to converge.) To establish which terms in the model represented significant effects, comparisons between the full model and nested models were made to check whether the inclusion of each term improved the fit of the model.

The full model represented a better fit to the data than one including fixed effects only of polarity, predictability and their interaction ($\chi^2(4) = 10.55$, $p = .032$), indicating that there was either a main effect of truth value or an interaction involving this variable. However, this was not the case for models omitting polarity ($\chi^2(4) = 5.50$, $p = .240$) or predictability ($\chi^2(4) = 6.20$, $p = .185$), suggesting that there were no significant main effects or interactions involving either of these factors.

The full model also represented a marginally better fit to the data than one including all terms except interactions involving truth value ($\chi^2(3) = 8.29$, $p = .040$), suggesting that truth value was involved in an interaction. However, as interactions involving the other two factors did not improve the model fit and the significance of this improvement was marginal, this was taken as an indication that any interaction involving truth value was a weak effect, and its primary effect on response accuracy was a main effect.

Coefficients representing the simple effects of each factor at each level of the other factors were computed based on the full model. These are listed in Table 2.2; each coefficient represents the difference in the log odds ratio of obtaining each outcome at the listed level of the factor and its reference level. The main effect of truth value indicated by the model comparisons described above was found to operate specifically in the case of low predictability, affirmative sentences. A weaker simple effect of predictability in the case of true, affirmative sentences was also identified. No other coefficients representing simple effects were significant.

2.3.2 Event-related potentials

Data preparation

Continuous EEG was re-referenced to the mean of the signal at M1 and M2 and segmented into epochs for analysis of ERPs. Based on previous findings indicating the region in which the N400 component is detected (e.g., Kutas & Federmeier,

Variable	β	z	p	95% CI	
				lower	upper
Falsity					
High pred.					
Aff.	-0.01	-0.04	.971	-0.40	0.38
Neg.	-0.03	-0.13	.898	-0.42	0.37
Low pred.					
Aff.	-0.67*	-3.23	.001	-1.08	-0.26
Neg.	-0.04	-0.22	.827	-0.44	0.35
High pred.					
True					
Aff.	-0.43*	-2.13	.033	-0.83	-0.03
Neg.	0.02	0.10	.922	-0.36	0.39
False					
Aff.	0.23	1.29	.197	-0.12	0.59
Neg.	0.04	0.19	.848	-0.33	0.41
Negation					
True					
High pred.	0.05	0.20	.772	-0.32	0.33
Low pred.	-0.39	-1.94	.052	-0.79	0.00
False					
High pred.	0.04	0.19	.847	-0.33	0.41
Low pred.	0.23	1.29	.197	-0.12	0.59

Table 2.2: EEG experiment: simple effects on response accuracy of each factor at each level of the other factors. For example, the first row provides the effect of falsity on the proportion of correct responses in a high predictability, affirmative condition. The reference levels of each factor are true (truth value), low (predictability), and affirmative (polarity).

2011), a region of interest was defined in the centroparietal area, consisting of channels C3, Cz, C4, CP1, Pz and CP2.

The onset of presentation of the critical word (the object name) in each critical (non-filler) trial was taken as the zero time point, and the signal was extracted for 200 ms preceding and 800 ms following this point, for each channel. In this window, the signal was corrected for artefacts arising from blinks using the automatic procedure available in the BESA software package (BESA 5.2, BESA GmbH; Berg & Scherg, 1994). Filler trials and trials with incorrect responses to the truth value judgement task were excluded from the analysis (12% of all critical trials); in total, 20,244 electrode-epochs from 3,374 trials were extracted from the recordings. All subsequent cleaning and analysis procedures, including rejection of noisy channels and movement artefacts, were carried out using R (R Core Team, 2018).

Each data point in each epoch was normalised to a baseline by subtracting the average magnitude of the signal in the 200 ms prior to the zero time point. To identify artefacts, epochs were binned by taking the average magnitude of the signal in each 20 ms window and comparing adjacent bins; epochs with more than a 20 μV difference between adjacent bins were rejected (5% of all extracted epochs). Epochs containing values falling outside the range of -30 to $30 \mu\text{V}$ were also rejected (2% of the remaining epochs). Next, based on visual inspection of the data for noisy or otherwise bad channels, data from a small number of channels were rejected wholesale for some participants (channels CP1, C3 and Pz were each rejected in the case of single participants; for a fourth participant, C4 and CP2 were both rejected). Additionally, another participant's data were rejected completely due to anomalous data across multiple channels. Following this procedure, 17,054 epochs from 3,149 trials were retained for analysis.

Data from the retained epochs were filtered using an order 3 Butterworth passband filter with a low pass cutoff value of 40 Hz and a high pass cutoff value of 0.1 Hz, and detrended by subtracting the overall linear trend of the epoch. Figure 2.4 illustrates the mean overall pattern seen at each electrode of interest over the timecourse of the epoch, for each condition.

To compare conditions in the time window of interest, the average magnitude of the signal between 300 and 400 ms after the onset of critical word presentation was computed for each epoch. These values were entered into the statistical analyses described below.

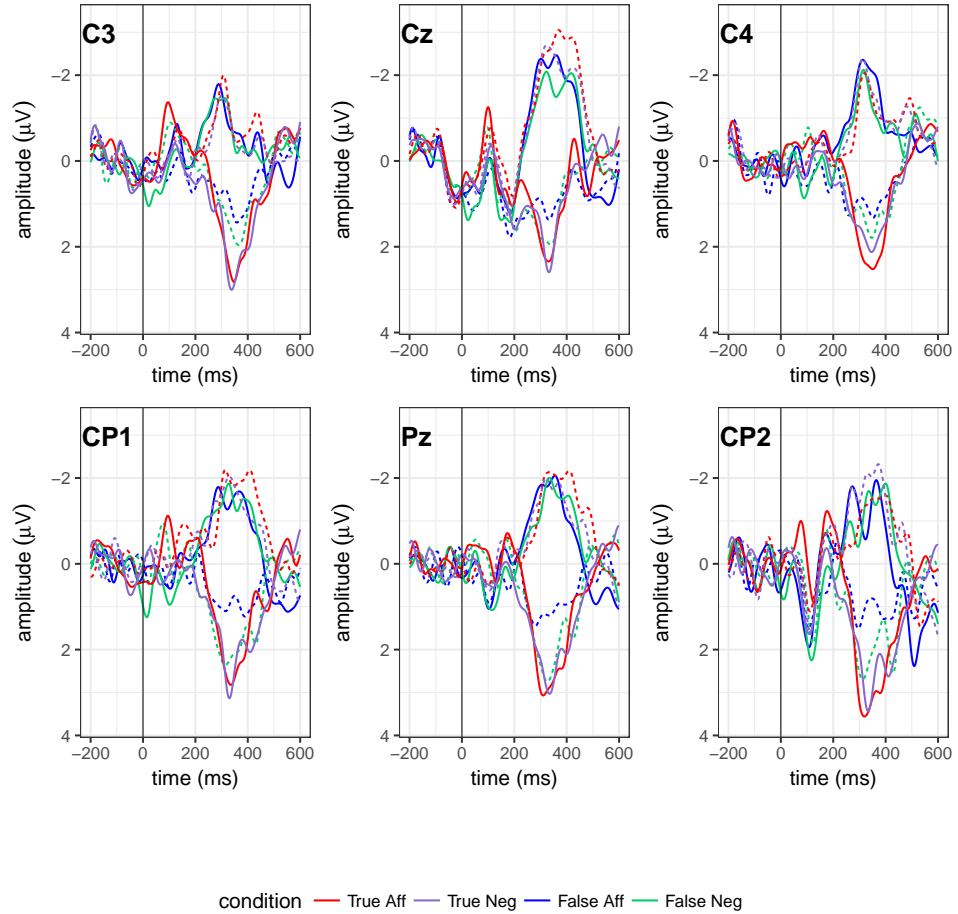


Figure 2.4: Mean timecourse of EEG in the window of interest: the mean signal recorded at each electrode, across all participants, by condition. The 0 ms time point represents the onset of presentation of the critical word (the name of the critical object in the sentence). Solid lines represent high predictability conditions; dotted lines, low predictability conditions.

Modelling of effects

To compare the magnitude of the N400 response across experimental conditions, a multi-level linear regression model was constructed over the average magnitude of the signal recorded on each trial, across all electrodes of interest, in the 300 to 400 ms time window, using the lmer function of the same R package used for accuracy data. A full model (formula 2.1), containing terms for the fixed main effects of each of these factors and their interactions (including the three-way interaction among all variables) and a random effect of truth value by participant (the maximal random effects structure that allowed the model to converge) was compared to models omitting the main effects and interactions of each factor in turn, using the same approach as for the behavioural data described above.

$$\text{amplitude} \sim \text{polarity} * \text{predictability} * \text{truth value} + (\text{truth value} | \text{participant}) \quad (2.1)$$

The full model represented a better fit to the data than the model without terms involving truth value ($\chi^2(4) = 295.6$, $p < .001$), indicating the presence of either a main effect of or an interaction involving truth value; the same was true for the model without predictability ($\chi^2(4) = 299.2$, $p < .001$). However, the full model did not represent a significantly better fit to the data than the model without terms involving polarity ($\chi^2(4) = 2.6$, $p = .631$), indicating that there was no main effect of polarity and that this factor was not involved in any interactions.

To test specifically for an interaction between truth value and predictability, the full model was compared to a model with only a fixed effect of each of these variables. The full model represented a better fit to the data in the case of truth value ($\chi^2(3) = 293.1$, $p < .001$) and predictability ($\chi^2(3) = 292.7$, $p < .001$), indicating the presence of an interaction between these factors.

Following these tests, coefficients were estimated for the full model to examine the simple effects of each factor at each level of the other factors. As shown in Table 2.3, these indicated that false sentences elicited a larger N400 ($M = -1.57 \mu\text{V}$, $SD = 5.83$) compared to true sentences ($M = 2.22 \mu\text{V}$, $SD = 6.17$) in the case of high predictability sentences, whereas (in a surprising finding, especially as it was the case for affirmatives as well as negatives), true sentences elicited a larger N400 ($M = -1.78 \mu\text{V}$, $SD = 5.28$) compared to false sentences ($M = 1.41 \mu\text{V}$, $SD = 6.31$) in the case of low predictability sentences. These findings are illustrated in Figure 2.5.

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Falsity						
High pred.						
Aff.	-4.01*	-9.92	380.37	< .001	-4.87	-3.26
Neg.	-3.64*	-8.94	372.58	< .001	-4.44	-2.84
Low pred.						
Aff.	2.90*	7.09	379.39	< .001	2.10	3.70
Neg.	3.49*	8.51	381.89	< .001	2.69	4.30
High pred.						
True						
Aff.	4.14*	10.38	3100.45	< .001	3.36	4.92
Neg.	3.89*	9.67	3102.88	< .001	3.11	4.68
False						
Aff.	-2.83*	-6.92	3102.76	< .001	-3.63	-2.03
Neg.	-3.24*	-8.02	3105.07	< .001	-4.03	-2.45
Negation						
True						
High pred.	-0.26	-0.65	3102.25	.518	-1.05	0.53
Low pred.	-0.02	-0.04	3098.54	.968	-0.80	0.77
False						
High pred.	0.17	0.42	3104.99	.674	-0.62	0.96
Low pred.	0.57	1.40	3103.33	.161	-0.23	1.38

Table 2.3: EEG experiment: simple effects on N400 amplitude of each factor at each level of the other factors. For example, the first row provides the effect of falsity on the amplitude of the N400 in a high predictability, affirmative condition.

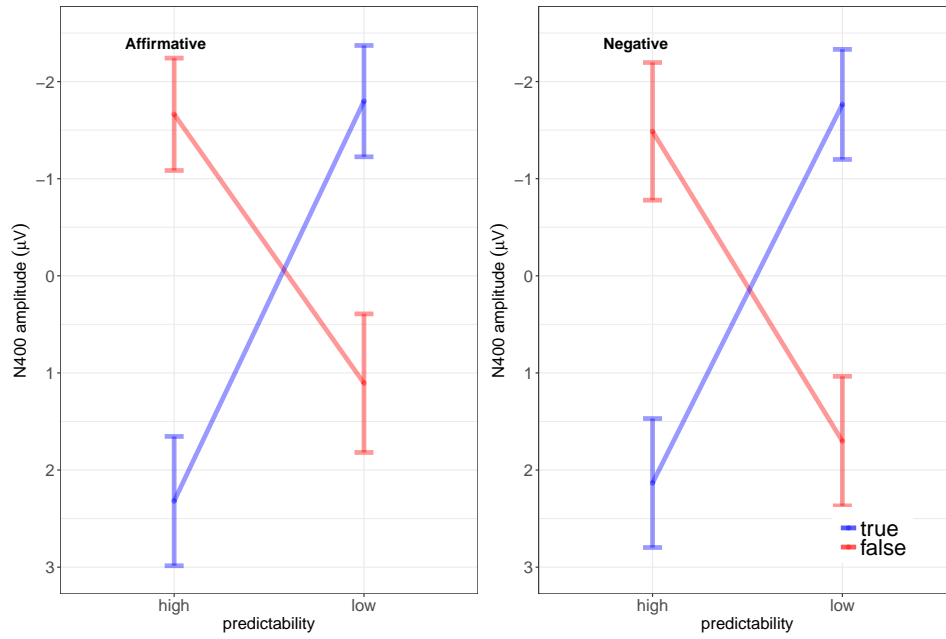


Figure 2.5: EEG experiment: mean N400 amplitudes in each condition. Error bars represent standard errors.

2.3.3 Memory test

Participants performed around or slightly above chance across all conditions (Table 2.4). The relevant coefficient of a logistic regression model over the proportion of correct responses, including fixed and random effects of polarity, suggested that there was no significant effect of polarity, $\beta = -0.00$, $z = -0.00$, $p = .997$, 95% CI = $[-0.49, 0.49]$.

Truth value	Polarity	Familiarity	Proportion correct	
			<i>M</i>	<i>SD</i>
True	Affirmative	Old	0.54	0.19
True	Affirmative	New	0.49	0.23
True	Negative	Old	0.49	0.24
True	Negative	New	0.68	0.19
False	Affirmative	Old	0.56	0.18
False	Affirmative	New	0.66	0.22
False	Negative	Old	0.55	0.22
False	Negative	New	0.63	0.21

Table 2.4: EEG experiment memory test: proportion correct in each condition of the memory test following the main experiment.

2.4 Discussion

This experiment examined the effects of manipulating the predictability of a critical word in a sentence on processing of affirmative and negative sentences, which could be true or false, while presenting only pragmatically felicitous sentences in a highly constrained, episodically-manipulated context. The results of a behavioural truth value judgement task showed that participants were approximately equally good at judging whether these sentences were true or false, with slightly worse performance for false sentences in a small subset of conditions. However, the N400 responses evoked by the critical words in the sentences varied greatly and did not conform to the original hypotheses.

Contrary to hypothesis 1, there was no main effect of truth value. Instead, a strong interaction between truth value and predictability was observed, with true sentences eliciting a reduced N400 compared to false sentences in the case of high predictability and an enhanced N400 in the case of low predictability. This was the case regardless of polarity, contrary to hypotheses 2 (that truth value would interact with polarity) and 3 (that this interaction would operate specifically in the case of low predictability, producing a three-way interaction). Although the findings were not in line with hypothesis 3, it is possible that this was a result of insufficient power to detect the three-way interaction, as the conditions did approximately follow the hypothesised pattern, but the interaction did not reach significance in the model comparisons.

The pattern of effects is readily interpretable in the case of the highly predictable conditions. Here, a critical word that met the strong prediction that could be generated elicited a reduced N400, whereas one that violated the strong predic-

tion elicited a large N400. This was the case regardless of whether the sentence was affirmative or negative, representing a replication of previous findings that negation can be interpreted incrementally, producing the same results as affirmatives, when it is pragmatically felicitous. Here, the pragmatic felicity of the sentences may have been further enhanced by the constrained episodic contexts meaning that there was only a single possibility to complete the predictable sentences truthfully. However, the pattern observed for low-predictability sentences is more difficult to interpret and may indicate that the interpretation just described does not capture the effects actually driving the pattern obtained.

If the reversal of the N400 effect between low- and high-predictability sentences had been observed only in the case of negative sentences (as in the original hypotheses), this would have been a strong indication that participants were not able to interpret negation incrementally online in a low-predictability condition, leading them to make erroneous predictions about the critical word even though the sentences were pragmatically licensed. However, this reversal was also seen in the case of affirmative sentences, where it cannot be explained by a failure to incrementally incorporate information. Therefore, it is most likely that this reversal can be attributed to some other factor common to both affirmative and negative sentences. The effect is particularly surprising because even low-predictability conditions featured a rather predictable critical word (with only two available possibilities) compared to most sentences encountered in natural language.

There are two main and related possible explanations for the overall pattern of findings. Both relate to the fact that all the conditions could also be categorised according to whether the object mentioned in the sentence (i.e., the critical word) was the unique object (that is, the action that succeeded for this object failed for the other two objects in the animation) or one of the non-unique objects (that is, the action that succeeded for this object also succeeded for a second object in the animation). This is illustrated in Figure 2.6. In the case of both affirmatives and negatives, the true, high predictability condition and the false, low predictability condition involved mentioning the unique object; these were also the same conditions that elicited a reduced N400, whereas the remaining four conditions, in which a non-unique object was mentioned, elicited a large N400.

This possible confound could have given rise to the overall pattern in two different ways. First, the simple salience of the unique object within the animation may have been sufficient to activate this concept to a greater extent than either of the non-unique objects, meaning that the unique object was primed simply by

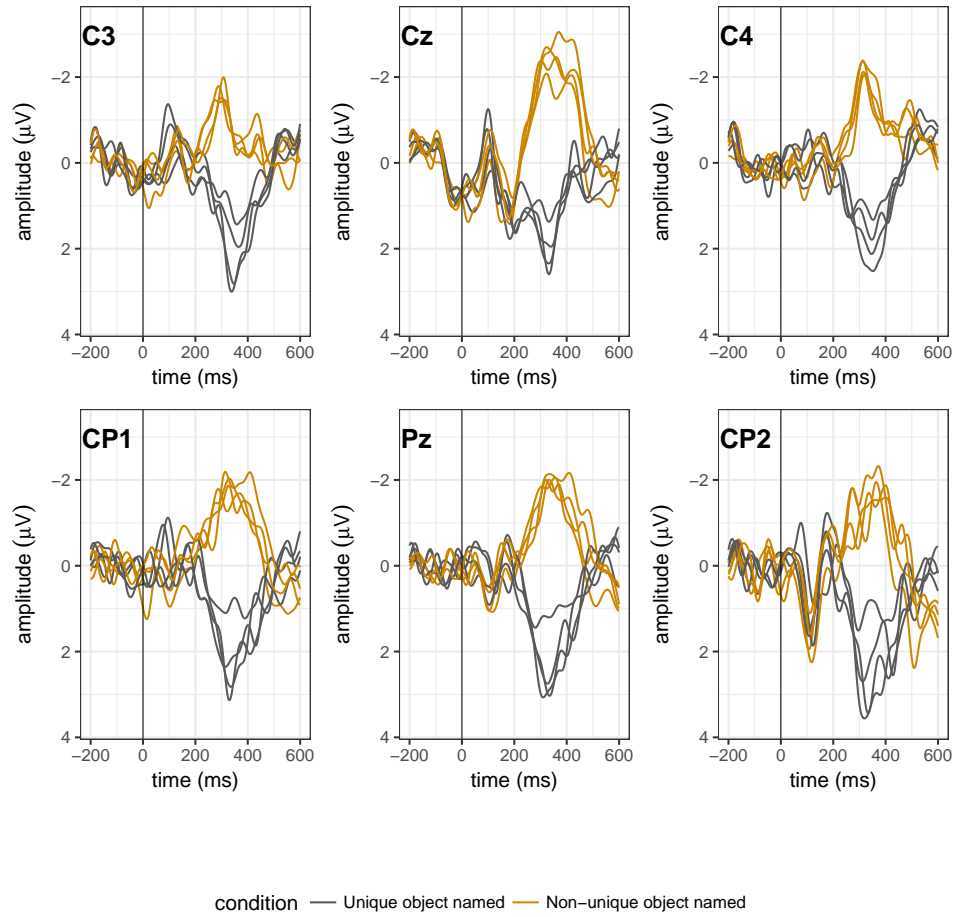


Figure 2.6: Re-visualisation of mean timecourse of EEG: this figure shows the same data as Figure 2.4, with conditions colour-coded according to whether the object mentioned in the condition is the unique object (the one sharing an action outcome with no other object) or one of the non-unique objects, rather than according to experimental condition. This visualisation illustrates the fact that this aspect of how the conditions were constructed may account for the overall pattern of effects.

virtue of participants' paying more attention to this aspect of animation. Thus, regardless of the sentence's meaning or truth value, a mention of the primed object elicited a reduced N400.

Alternatively, participants may have employed a deliberate or unconscious strategy in completing the task that exploited the uniqueness of one of the objects as a memory aid. The truth value judgement task was not an easy one, as the animation proceeded quickly and in order to make a correct judgement about the sentence, the participant needed to remember associations between three different objects and two different actions. However, because only actions and objects that had in fact featured in the animation could be mentioned in the sentence, an available strategy to reduce this memory load would have been to ensure that a single relationship was memorised for each animation: the combination of the unique object and the action that was applied to it. This information would enable the participant to judge any sentence true or false (in the case of critical trials). Thus, participants had good reason to focus on the unique object (enhancing the potential priming effect based on its salience) and, furthermore, may have formulated their predictions about upcoming material as (for example) either "*leaf*" or "anything but *leaf*". Thus, although the participant may have been making a correct prediction in this scenario, either case may have enhanced activation of the unique object, even when the prediction was effectively "not the unique object". The presence of fillers, in which there was often no unique object, may have discouraged the use of this strategy to an extent; however, fillers were only about half as numerous as critical trials, and even if participants could not use this strategy in the case of fillers, it remained available to them on the majority of trials. If participants did use this strategy, it does appear that they were able to do so equally effectively for affirmatives and negations; however, it is difficult to interpret the findings as concrete evidence that processing proceeded equally incrementally in both cases, because the first account suggested here (in which the pattern emerged purely as a result of the salience of the unique object) does not require any sentence interpretation to have occurred at all in order to obtain the pattern of results observed.

One piece of evidence suggesting that the former interpretation may be more likely than the latter (which relies on the participant's use of a particular strategy, beyond the possibly automatic effects of salience) is the fact that encountering a negation of a concept (as in "not *leaf*" or "anything but *leaf*") does not activate the mental representation of that concept to as great an extent as the same concept

presented with negation elsewhere in the sentence. For example, MacDonald and Just (1989) found that presenting negated concepts such as “not bread” primed related probes (in this case, “butter”) to a reduced extent in comparison to non-negated concepts (“bread”). This is in line with the view that negation should be interpreted as an instruction to shift one’s attention away from the negated entity or proposition and towards another (e.g., De Mey, 1972). Therefore, the use of a strategy in which a participant formulates the prediction “not *leaf*” or “anything but *leaf*” might not be expected to produce as complete a reversal of the effects between the high-predictability and low-predictability conditions as seen here.

Another concern for the interpretation of the results might be how the independent variables interacted with one another in determining whether a sentence could be regarded as “predictable” or not. The condition into which each sentence fell was designated on the basis that the participant would make predictions about upcoming material using the assumption that the sentence would be true. This is a reasonable Gricean assumption to make during everyday use of language (as pointed out in the context of negation processing by Tian & Breheny, 2015), but in the context of a series of artificially-constructed sentences that were equally as likely to be false as true, it is possible that this designation may not have been completely valid: that is, participants may have been unable to formulate strong predictions about some sentences designated highly predictable, and vice versa. The predictability of sentences becomes further difficult to interpret if participants did in fact have difficulty incrementally incorporating negation into their representations, and thus into their predictions. If negation was initially not taken into account, participants could have formulated an erroneously strong prediction in the case of unpredictable sentences, and vice versa. For instance, in the case of the scenario illustrated in Figure 2.1, the negative sentence “The fairy didn’t deposit a square around the...” could have two possible completions to make it true (*foot* or *blanket*), meaning that a participant who had taken account of the presence of negation incrementally would generate a weak prediction. However, a participant who had not done so (and was therefore focusing on objects for which the “deposit a square” action had succeeded, rather than failed) would make an erroneously strong prediction of *leaf*. Their predictions would then turn out to be valid in the case of a true sentence and invalid in the case of a false sentence, as intended by the experimental design; however, the reversal of predictability in these cases increases the difficulty of interpreting the findings.

The fact that participants were explicitly asked to evaluate the truth value of the

sentences in the behavioural task may have further enhanced some of these issues, for example by increasing their focus on considering which objects were paired with which action to a greater extent than would be the case in a more natural setting. The operation of a “truth-evaluation mindset” for sentence processing could also in itself have influenced participants’ cognition during the task. This type of effect has been shown by Wiswede, Koranyi, Müller, Langner, and Rothermund (2013), who identified a late negative-going ERP component, specific to false affirmative sentences, that was present only in a group of participants in which such a mindset was induced by the use of a truth value judgement task, compared to a sentence-matching task. Although this component is distinct from the N400, the fact that it can be identified (significantly prior to the actual response, which was delayed for 1,500 ms) suggests that this type of mindset may meaningfully impact how sentences to be evaluated are processed online, beyond the need to provide a judgement after processing.

Overall, the pattern of results indicates that the experimental manipulations were unlikely to have been the main drivers of the effects observed, because in this case, the N400s evoked by affirmative, unpredictable sentences are very difficult to explain. The most likely explanation is that the N400 was mainly influenced by the salience of the unique object in the animations associated with critical trials. This means that although the findings provide no evidence that predictability affects the incremental processing of negative, pragmatically felicitous sentences, they also cannot be interpreted as concrete evidence that this is not the case (i.e., that even low-predictability negations are processed as readily as affirmatives given adequate pragmatic licensing). An alternative paradigm is required to tackle this question further, either by avoiding the confounding effects of uniqueness (a difficult proposition when it is this factor that permits manipulation of the predictability of the sentence) or by unpicking whether this is in fact the relevant generator of these effects or can be discounted. One possibility for the latter approach would be to test multiple levels of predictability: for example, by including a condition in which there are three alternatives for the critical word, as well as one and two. If predictability exerts much more influence on processing in the shift from one to two available predictions than in the shift from two to three, this could be an indication that the special status of a “unique” object is driving the effects.

Chapter 3

Mousetracking Experiment 1: Truth Value Judgement

3.1 Introduction

3.1.1 Negation and predictability

It has long been clear, as discussed in the General Introduction and in the preceding chapter, that sentences containing negation appear to impose extra processing costs compared to affirmatives. Various studies (e.g., Fischler et al., 1983; Kaup & Lüdtke, 2007; Tian et al., 2010) have suggested that this cost may take the form of impairment to the usually incremental nature of processing, in which new information is interpreted online and predictions for upcoming material updated continuously (Altmann & Mirković, 2009; Rayner & Clifton, 2009).

However, more recently it has emerged that, given the right conditions, this processing disparity between negatives and affirmatives can be mitigated or even erased. In particular, Nieuwland and Kuperberg (2008) demonstrate that, when negated sentences are presented in a context that makes them fully pragmatically licensed, the distinction between affirmatives and negatives (as indexed in this case by the N400 component of the event-related potential to a critical word) disappears. Similar results have been obtained using a mouse-tracking method (see below) by Dale and Duran (2011).

Nieuwland and Kuperberg (2008) draw the conclusion that earlier studies (e.g., Fischler et al., 1983) have misinterpreted the nature of negation processing by unfairly presenting infelicitous negations for comparison with more felicitous affirmatives. It is generally implicit in the formation of a negation that some

state of affairs under discussion requires denial or contradiction (see General Introduction), otherwise there would be no Gricean reason to communicate a negative; for instance, doing so would be less than maximally informative (Grice, 1975). Thus, the pragmatic context required to license a negative statement is much more constrained than the context required to license an affirmative, and to a corresponding extent, laboratory studies not accounting for this disparity have tended to compare affirmatives and negatives differing in the extent to which they are felicitous.

However, this confound cannot account for all the evidence comparing affirmatives and negatives. If this were the case, pragmatically licensed affirmatives and negatives should be equally easy to process (as in the above-mentioned studies), but poorly licensed affirmatives and negatives should also be equally *difficult* to process; this is inconsistent with Nieuwland and Kuperberg (2008), who report an interaction between polarity and pragmatic felicity, with a larger effect of infelicity on negatives. Therefore, it seems that the effect cannot simply be attributed to the fact that affirmative sentences are generally less heavily unlicensed when presented outside an appropriate pragmatic context; instead, there is something about pragmatic infelicity that interacts with negation to produce a stronger impact on processing.

As discussed in Chapter 2, one aspect of pragmatic infelicity to which this effect could potentially be attributed is a lack of predictability. Although predictability and pragmatic felicity are related, they are not reducible to one another; predictability is simply a component or common side-effect of strong felicity. Therefore, the hypothesis that predictability is the “active ingredient” in pragmatic felicity with respect to its interaction with polarity can be tested by manipulating the former while holding the latter constant.

One way of doing this is through the use of contexts that produce episodic associations between concepts or entities, inducing comprehenders to make predictions on the basis of these local, temporary associations rather than on the basis of broad semantic knowledge (which is more commonly relied upon in this type of manipulation, as discussed in Chapter 2). In the presence of a visual stimulus, utterances describing the image may be highly felicitous (because the presence of the image places its contents or characteristics in the domain of discourse), and, separately, contain highly predictable or less predictable elements (depending, for example, on how many entities in the image match a particular description). For instance, the sentence “The top shelf contains a candle” may have a highly

predictable final word (if the candle is the only item on the top shelf, in the visual context) or a less predictable one (if multiple items meet this criterion), but is no more or less felicitous in either case.

There are several advantages to this approach to manipulating predictability. First, this is a tightly controllable manipulation, because the number of candidates for the critical word in such a sentence can be fixed precisely by the content of the visual context. Second, and relatedly, the participant's ability to make a prediction does not rely on their world knowledge or pre-existing associations, which may vary between participants, vary in strength, or differ from the researcher's expectations in unpredictable and undetectable ways. Finally, the problem of interference from semantic priming of associated concepts is avoided.

For these reasons, the present experiment investigated the relationship between predictability and polarity in their impact on online sentence processing, through the use of sentences presented in episodic visual contexts.

3.1.2 The mouse-tracking methodology

As outlined in the preceding chapter, EEG (and specifically, in this case, the N400 component of the ERP) provides extremely useful access to early cognitive processes, avoiding many of the pitfalls associated with behavioural methods. However, plenty of detailed information on participants' cognitive processing can also be accessed without directly measuring brain activity. Although many traditional behavioural measures (such as response accuracy or speed in completing a task) yield only offline measures, in which all cognitive processes underlying a task are represented by a single datapoint, online behavioural measures are also available to a greater or lesser degree. These purport to offer insight into participants' cognitive processing or mental states at intermediate stages during a task.

Computer mouse-tracking is one such methodology, which exploits the idea that when participants use a computer mouse or similar pointing device to respond to stimuli, the trajectory followed by the cursor represents aspects of their cognition while formulating and executing stages of the response. This notion is based on the underlying principle that there is a close link between cognition and action. The strongest form of this principle draws on theories of embodied cognition (see, e.g., Barsalou, 2010), although the notion that motor actions can provide access to information about cognition does not necessarily rely on this view, especially when cognition is ongoing and updated during the process of making a motor response.

Use of a computer mouse is not only routine and intuitive for participants, but also somewhat automatic, in that very little cognitive overhead is involved in operating the mouse and participants perceive almost no task demands relating to this method of responding. This is in contrast to some other measures which purport to give similarly detailed levels of information, such as the signal-to-respond paradigm (Kent, Guest, Adelman, & Lamberts, 2014), which requires intensive training to allow participants to respond when cued to do so, and even after such training imposes a high level of demand in addition to the task of interest. Additionally, mouse movements are very cheap and easy to record (Freeman & Ambady, 2010, provide a simple interface for presenting stimuli and recording data in this paradigm) and there are no requirements for set-up overhead time, technical skills, or equipment beyond a standard computer and mouse, enabling quick and easy collection of potentially large amounts of data.

As a result of these advantages of the methodology, its use is becoming increasingly widespread in several areas of cognitive psychology, including language research. Kent, Taylor, Taylor, and Darley (2017) review the theory underlying the mouse-tracking approach and its application to various topics of interest in memory and language, including categorisation (Dale, Kehoe, & Spivey, 2007), decision-making (McKinstry, Dale, & Spivey, 2008), and two-step interpretation of scalar implicatures (Tomlinson et al., 2013).

3.1.3 Design and hypotheses

Mouse-tracking Experiment 1 aimed to analyse mouse trajectories collected during truth-value judgements of sentences varying in polarity and predictability, to assess whether the predictability of a critical word modulates the extent to which comprehenders can incorporate negation incrementally into their predictions. If participants are able to process incoming elements of a sentence online and formulate predictions for upcoming material on the basis of these incremental interpretations, this should facilitate performance in a truth-value judgement task: if their prediction is fulfilled by the critical word in the sentence, it should be relatively straightforward to make a TRUE judgement, and similarly, if it is contradicted, a FALSE judgement should be easy to arrive at. In contrast, if participants cannot make a useful prediction for the critical word, they should respond more hesitantly or erroneously; and if they have made an incorrect prediction (for example, by failing to incorporate incrementally the correct interpretation of a negating element), they may be initially drawn towards the foil response (TRUE for a false

sentence, and vice versa).

To test this, participants completed a task in which, on each trial, they made a truth value judgement of a sentence relating to an image presented immediately beforehand. Each image consisted of a 3×2 grid in which one of the rows was filled with three objects and the other could contain one, two, or three objects (see Figure 3.1 in the Methods section). Sentences were of the form “The top row contains / doesn’t contain the *lamp*”, meaning that their predictability could be manipulated by varying the number of objects in the row in question. Participants provided their judgements by moving the mouse from the bottom of the screen to TRUE and FALSE response boxes in the top left and right corners, respectively. Thus, a 2 (polarity: affirmative or negative sentence) $\times 3$ (predictability: high, medium, or low) design was employed for the independent variables of interest. Sentences could also be true or false.

The hypotheses for this experiment were based on the notion that predictability is the component of pragmatic felicity that disproportionately affects processing of negative sentences. However, because only pragmatically felicitous sentences were included, no main effect of polarity was expected on any of the dependent measures indexing processing. Predictability, in contrast, was expected to exert such a main effect, with ease of processing correlating with increased predictability. Most importantly, if (as hypothesised) predictability is an important component of infelicity in its impact on the processing of negations, polarity should be expected to interact with predictability. In particular, the detrimental effect of reducing predictability from high to medium and from medium to low was expected to have a stronger impact on participants’ performance in the case of negative sentences; furthermore, the increase in the proportion of mouse trajectories exhibiting active attraction towards the wrong answer as predictability was reduced was expected to be greater in the case of negations compared to affirmatives.

Assessing “anti-prediction”

Three levels of polarity were specifically compared in order to assess the impact of the possible “anti-prediction” strategy discussed in Chapter 2. This strategy would only be available to participants if they were making mistaken predictions as a result of failing to update on the basis of negating elements in a sentence. For example, if the participant (incorrectly) believes that the sentence could truthfully end with any of three different items (A, B, or C), but can *exclude* a fourth (D), a strategy to minimise resource consumption may be to predict “not D” rather

than “A, B, or C” (c.f. Orenes, Beltrán, & Santamaría, 2014). Doing so could cause activation of D in memory (even though it is precisely the opposite of the concept referred to), in turn causing initial attraction towards TRUE as a response upon hearing D, even though it is the converse of the prediction. This could affect the interpretation of results, because an initial attraction towards D would be taken in this paradigm as evidence of incremental processing, whereas it in fact results from the compounding of two “errors” (failing to process negation incrementally, and attraction towards the opposite of an erroneous prediction).

This strategy, if employed by participants, should only affect the high predictability condition (and then only in the case of non-incremental interpretation of negative sentences), because predictions like “not D or E” are much less likely to be effective in terms of resource consumption. Therefore, investigating three levels of predictability would allow a comparison between the impact of high and medium predictability in particular. If the difference in measures of processing ease for negations shows a much larger advantage of high over medium predictability than that of medium over low predictability, this might constitute a suggestion that participants’ use of an anti-prediction strategy was artificially responsible for their apparently enhanced performance in high predictability negations.

3.2 Methods

3.2.1 Materials

Stimuli consisted of images of highly-identifiable, easily-named, everyday inanimate objects, selected for consistent naming by a small sample of British English speakers. On each trial, images were presented in a grid of three columns and two rows, accompanied by a sentence presented auditorily, of the form “The top / bottom row contains / doesn’t contain the *balloon*”. Either the top or the bottom row of the grid always contained three images, while the other row could contain one, two, or three images. In this way, the predictability of the critical object name that would appear at the end of the sentence to make it true was manipulated. High-predictability sentences only had one possible ending that would make them true; for medium-predictability sentences, there were two possibilities; and for low-predictability sentences, there were three possibilities.

In total, 72 sets of visual stimuli were constructed, each consisting of 18 versions: six with four objects, six with five objects, and six with six objects. Within each of these subsets, the critical object exchanged roles with one of the objects

in the other row in half the versions; and within each of these three subsets, the critical object appeared in the left, centre, or middle column of the grid. The variable number of objects appeared on the top row in half the image sets and on the bottom row in the other half. Each of the 18 versions within each set could be associated with four possible sentences: affirmative or negative, and true or false. Thus, each set of images was associated with $18 \times 4 = 72$ possible trials, of which half were critical trials and half fillers. Critical trials (design illustrated in Table 3.1) were equally distributed across conditions in a 2 (polarity: affirmative or negative sentence) \times 3 (predictability: high, medium, or low) \times 2 (truth value: true or false sentence) design. Filler trials did not fit into this scheme and were therefore not of experimental interest; there were always three possible candidates for a critical word that would make the sentence true, meaning that predictability was not manipulated. They were included so that participants would not alter their focus of attention or their strategy based on the knowledge that (on trials with fewer than the full 6 objects) those objects on the row that was not completely filled would always be the candidates for the critical word that would make the sentence true. Each participant completed a total of 216 of the $72 \times 3 = 216$ trials constructed, divided evenly across critical conditions and their equivalents among the fillers.

Audio recordings of the required sentences were prepared using the NaturalReader software (NaturalReader Version 11, NaturalSoft Ltd, 2015), employing a female British English speaker's voice with natural-sounding prosody. The audio files were manipulated using the Audacity(R) recording and editing software (Audacity 2.1.2, Audacity Team, 1999–2016), in order to homogenise the time elapsing from the onset of the sentence fragment to the onset of the critical word that would allow the participant to distinguish an affirmative and a negative sentence (i.e., *contains* or *doesn't*).

Trials were presented using the MouseTracker software package (Version 2.82, Freeman & Ambady, 2010), on a 60 cm monitor in which the grid of images, when visible, occupied an area of the centre of the screen measuring approximately 18 \times 12 cm. Response buttons, consisting of black rectangles overlaid with white text reading TRUE and FALSE, were located in the top-left and top-right corners of the screen, respectively.

In addition to the main set of trials, “catch trials” were included to encourage participants to pay careful attention to the objects in each grid. These consisted of an on-screen prompt, following a trial, to recall the objects that had been presented










Grid	Sentence			
	The bottom row contains the basket	The top row doesn't contain the basket	The bottom row contains the lamp	The top row doesn't contain the lamp
	True Aff, High Pred	True Neg, High Pred	False Aff, High Pred	False Neg, High Pred
				
				
	True Aff, Med. Pred	True Neg, Med. Pred	False Aff, Med. Pred	False Neg, Med. Pred
				
				
	True Aff, Low Pred	True Neg, Low Pred	False Aff, Low Pred	False Neg, Low Pred
				
				

Table 3.1: Mouse-tracking Experiment 1 example stimulus sets. Critical trials only are shown, using 36 of the 72 possible trials based on one set of images. The other 36 (fillers) use the “missing” sentences (e.g. “The bottom row doesn’t contain the basket/lamp”).

in the grid for that trial, and a pre-printed grid on paper for use in completing this task.

3.2.2 Procedure

Participants were seated in a quiet lab at a comfortable distance from the monitor. They were tested individually, with zero, one, or two other participants present, in sessions that lasted approximately 60 minutes.

The experiment began with a verbal explanation of the task from the experimenter, accompanied by practice trials, including a “catch trial” practice. Participants were asked to look carefully at every item in the grid on each trial, and to focus on clicking on the correct response button (“TRUE” or “FALSE”). They were informed that on catch trials, they should try to fill in the name of every object that had been present in the grid in the correct location; however, if they remembered that a particular location had contained an object but not the identity of the object, they could indicate this with an X. After four practice trials, each participant completed eight blocks, each consisting of 27 randomly ordered trials. They were encouraged to take a break between blocks. Catch trials appeared randomly throughout the experiment (30 per participant, with 10 at each level of predictability).

Each trial began with a screen containing a “START” button in the bottom centre location; participants were required to click here to initiate each trial, so that the mouse and cursor were reset to the same starting position. The visual stimulus (i.e., the grid of images) was then displayed in the centre of the screen for 3,000 ms, plus a further 1,500 ms for every additional object above 4 (i.e., 4,500 ms for 5-object trials and 6,000 for 6-object trials). After this time had elapsed, the grid disappeared and auditory presentation of the sentence (through stereo headphones) began. At the same time as the onset of the audio, the response options (“TRUE” and “FALSE”) appeared in the top-left and top-right corners of the screen, respectively, and the cursor was released to allow participants to complete their response when ready. Participants were required to initiate their response (by moving the mouse) within 5,000 ms; if they failed to do so, they received a warning message (“Please start moving as soon as you’ve finished, even if you are not fully certain of a response yet”). Mouse coordinates were sampled online every 32 ms during the response phase of each trial, based on a virtual coordinate space ranging from -1 to 1 in the x axis and 0 to 1.5 in the y axis, with the origin at the horizontal centre of the bottom of the screen. Response initiation and completion

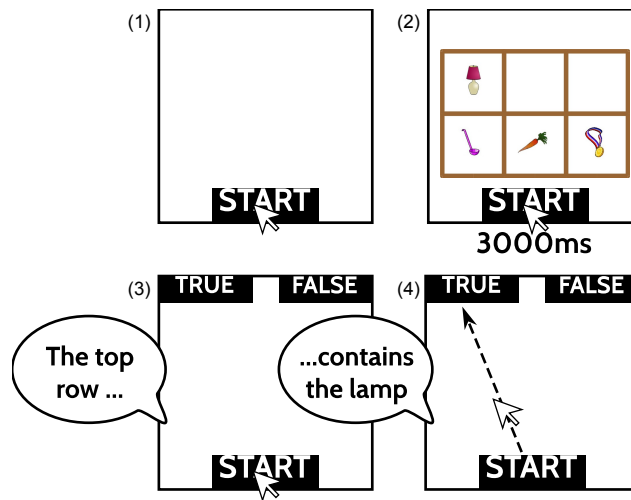


Figure 3.1: Mouse-tracking Experiment 1 example trial. Participant clicks the START button; visual stimulus is displayed; sentence is played and response options appear; participant makes their selection using the mouse. Stimuli and response buttons are for illustration and not to scale.

times and accuracies were also collected. After completing their response, the participant received feedback in the form of a green O (correct) or red X (incorrect) displayed for 300 ms. For trials accompanied by a catch trial, the memory test prompt appeared after the feedback; there was no time limit for responses to this part of the task. Figure 3.1 illustrates an example trial.

3.2.3 Participants

A total of 24 participants (21 female, aged 18–24 years [$M = 19.5$, $SD = 1.7$]) were recruited from in and around the University of Bristol community. All were native speakers of English, and all but two (Thailand, the Netherlands) had grown up primarily in the United Kingdom. Most were monolingual, but four had a second language of a good, advanced, or fluent standard (Thai, Welsh, or Dutch).

Informed consent was obtained from all participants prior to their participation, and they were compensated for their time with course credit or a small payment. Ethical approval for the experiment was granted by the University of Bristol's Faculty of Science Research Ethics committee (ethical approval code 31441).

3.3 Results

3.3.1 Data preparation and analysis

Across all participants ($N = 24$), the range of error rates was 3% to 22% ($M = 9\%$). Trials with incorrect responses were discarded from the analyses (i.e., 9% of all trials). Those trials with response completion times longer than 6,000 ms, and trials on which the participant took longer than 4,000 ms to initiate their response, were also excluded (a further 3% of the remaining trials). These thresholds were selected on the basis of visual examination of the distributions of response initiation and completion times.

For ease of comparison across all trials, data from trials with the correct answer “TRUE” (i.e., for which the target response was located on the left side of the screen) were reflected (all x-coordinates were sign-reversed) using the MouseTracker software package’s inbuilt analysis features (Freeman & Ambady, 2010).

The dependent measures analysed were response accuracy, response initiation time (measured from the time of cursor release to the onset of the participant’s mouse movement) and completion time (measured from the same starting point to the participant’s mouse click on the target response), and a proxy for trajectory shape (see below).

Data are presented below predominantly in the form of “violin” plots (produced using the ggplot2 package in R; Wickham, 2016), which illustrate the overall shape of the data across different conditions, allowing for visual comparison not only of means but also of the different degrees of skewness across conditions, and so forth.

Analysis of trajectory shapes

To characterise the shape of a path followed by the cursor from the starting position to the target response button, generally useful metrics are the “area under the curve” (AUC) and the “maximum deviation” (MD).

The former represents the total area of the screen lying between an ideal straight-line trajectory drawn between the starting and finishing points and the actual trajectory taken, and the latter the distance, at the furthest point as measured by an orthogonal line, between this ideal straight-line trajectory and the actual trajectory (see Figure 3.2). A greater AUC and MD may each represent a trial with a higher degree of curvature towards the foil response (and hence a degree

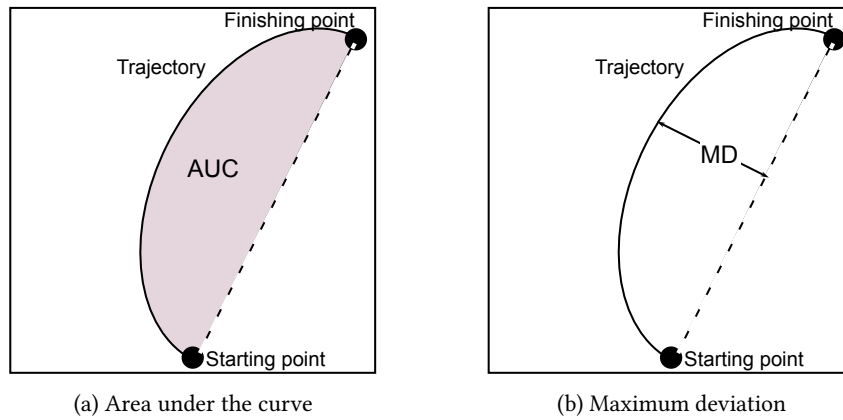


Figure 3.2: The calculation of area under the curve and maximum deviation as characterisations of a path. The solid line between the starting and finishing points represents the actual trajectory (in normalised time); the dotted line represents a straight line between the starting and finishing points.

of attraction to the foil can be inferred), although they provide slightly different characterisations: for example, a path where the participant wanders slightly around the screen throughout the trial could have a relatively low MD but a somewhat high AUC.

Both metrics were computed using the MouseTracker software package's inbuilt analysis features. The package employs a time-normalised version of the data for this function, in which all samples from each trial are divided into 101 bins of equal duration (regardless of the duration of the full trial) and the average x- and y-coordinates during each bin used to represent the location of the cursor during that time bin.

Because these measures typically fall into a bimodal distribution across a dataset of this type (including in the present case, as shown below), it is difficult to apply standard analysis techniques. Furthermore, neither metric necessarily captures the most pertinent characteristics of a given trajectory, or of a set of trajectories produced in a given task, as they reduce a rich set of time-course based data to a single statistic. Therefore, in addition to these measures, a clustering approach was used as a data-driven method of assessing whether trials exhibited attraction to the foil response. This approach was based on the notion (supported by bimodality observed across multiple dependent variables) that trajectories produced by participants fell broadly into two main categories: those in which the mouse was moved rather quickly and directly to the correct response option, and

those exhibiting some degree of attraction towards the wrong answer, with the latter type also tending to be initiated and completed more slowly. Categorising trajectories into these two clusters, which could differ between participants in their typical shape, could offer a way to identify the proportion of trials in each condition that exhibited some level of attraction towards the incorrect response option.

Trajectories were assigned to clusters based on their shape using the Hartigan and Wong (1979) implementation of the k-means algorithm with $k = 2$. Trials were first separated by participant so that clustering could be performed separately for each participant, and were represented as a vector of x and y coordinates at each timestep. (For trials shorter than the maximum response time cut-off, timesteps following the end of the trial were considered to have the same coordinates as the final timestep of the trial.) The algorithm randomly selected k vectors to represent the initial cluster centres, allocated all vectors to the nearest of these (such that the squared Euclidean distance was minimised), recalculated the cluster centres by taking the mean vector within each new cluster, and iterated this process either a maximum of 1,000 times or until convergence. The best outcome resulting from 1,000 different random starting configurations was selected, with a seed for the whole analysis set at 1. After this process was complete for every participant, the data were re-integrated for analysis of cluster allocation, so that a given participant's trial cluster represented by the cluster centre most similar to a straight line was treated as equivalent to that of other participants.

Statistical modelling of effects

Response accuracy, response initiation and completion time, and trajectory cluster allocation were each modelled separately to analyse the effects of the independent variables on these measures, using the R package lme4 (Bates et al., 2015). For proportional data (i.e., response accuracy and cluster allocation), the glmer function was used to examine a mixed effects logistic regression model with a logit link function; for continuous data (i.e., initiation and completion time), the lmer function was used to examine a mixed effects linear regression model. The independent variables of interest were sentence polarity (affirmative or negative) and predictability of the final word in the sentence (high, medium, or low, based on how many objects fitting the criteria at the beginning of the sentence were present in the image).

Although truth value of the sentence was also manipulated, this was for the

purpose of creating a viable task, rather than because this was also a variable of interest. Because true and false sentences were evenly distributed across the other conditions, ideally truth value would be ignored in the analysis, with any main effect applying equally across conditions. However, a problem with this approach is that truth value might also interact with the variables of interest. In particular, there is reason to expect an interaction between truth value and polarity (as observed, for example, by Gough, 1965; Wason, 1959, 1961), perhaps because of an intrinsic association between negation and falsity. In addition, although there is less of an a priori reason to expect that predictability might interact with truth value, or that a three-way interaction among all three variables might emerge, this cannot be ruled out. Thus, the effects of truth value might not be equal across all conditions in this design. Therefore, truth value was entered into the analysis alongside predictability and polarity, plus terms representing the interactions among these three variables, so that its effects could be taken into account.

For each dependent variable, a full model was constructed including fixed factors for all the above effects, as well as the maximal random effects structure that allowed the model to converge. Formula 3.1 exemplifies the full model using response time as the dependent variable. Here, the maximal possible random effects structure is given for illustration purposes, although this structure did not reach convergence for any of the dependent variables.

$$\begin{aligned} \text{response time} \sim & \text{polarity} * \text{predictability} * \text{truth value} \\ & + (\text{polarity} * \text{predictability} * \text{truth value} | \text{participant}) \end{aligned} \quad (3.1)$$

Next, this full model was compared to various nested models to investigate the presence of main effects and interactions, using the following procedure:

1. The full model was compared to models including only the fixed effects of, and interaction between, each of only two factors: e.g., polarity \times predictability only. If the former represented a significantly better fit to the data, this was taken as an indication of either a main effect of the third factor, or its involvement in an interaction.
2. For each factor for which this was the case, its involvement in an interaction specifically was tested by comparing the full model to a model including the fixed effects of, and interaction between, each of the other two factors, plus a fixed effect of the factor in question. If the full model represented a significantly

better fit to the data, this was taken as an indication that this factor was involved in an interaction with at least one other factor.

3. If the above tests indicated the presence of interactions, the full model was compared to a model including fixed effects of all three factors, plus each of the possible two-way interactions (polarity \times predictability, polarity \times truth value, and predictability \times truth value). If the former represented a significantly better fit to the data, this was taken as an indication of the presence of a three-way interaction.

Following these tests, the full model was used to estimate coefficients for the simple effects of each factor at each level of the other factors, and (where applicable) for the simple interactions of each pair of factors at each level of the third. Associated confidence intervals (using the Wald method) and p values (using the Satterthwaite approximation) were also computed.

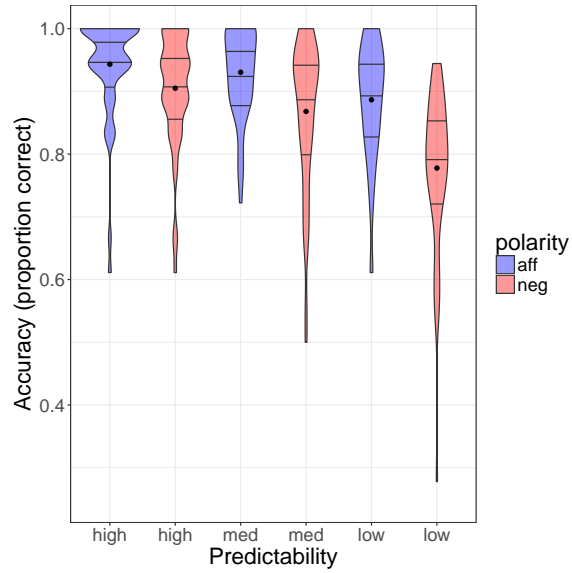
3.3.2 Response accuracy

Although participants found this task challenging according to their informal reports to the experimenter, they were generally able to complete the main task accurately, providing correct responses on the majority of trials (proportion correct: $M = 0.89$, $SD = 0.06$). Some conditions were more difficult than others, with a more skewed distribution of accuracies tending to appear for false sentences and negated sentences (Figure 3.3).

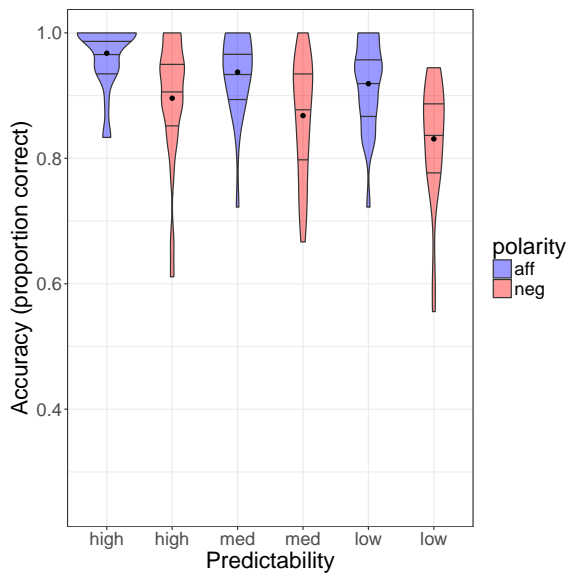
Figure 3.4 summarises the differences in response accuracy across all conditions. As described above, these differences were modelled across all conditions, with a full set of fixed factors and a random effect of polarity included by participant (models involving a more complex random effects structure failed to converge). This model is shown in formula 3.2.

$$\begin{aligned} \text{cbind(prop. correct, prop. incorrect)} &\sim \text{polarity} * \text{predictability} * \text{truth value} \\ &+ (\text{polarity} | \text{participant}), \text{family} = \text{binomial}(\text{link} = \text{"logit"}) \end{aligned} \quad (3.2)$$

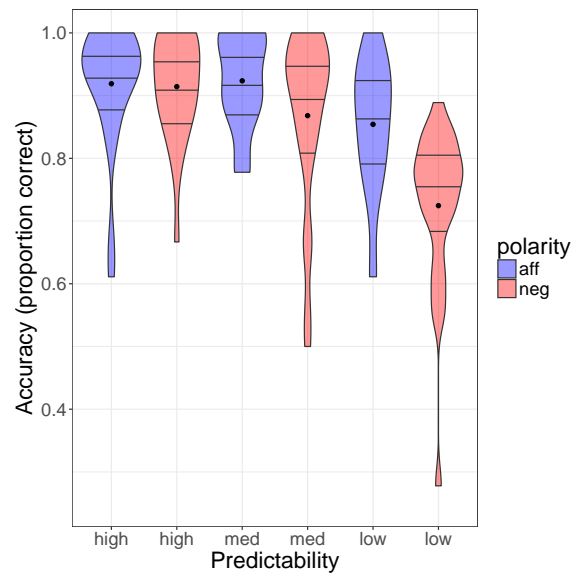
Model comparisons testing for the presence of main effects and/or interactions were significant for polarity ($\chi^2(6) = 36.89$, $p < .001$), predictability ($\chi^2(8) = 95.78$, $p < .001$), and truth value ($\chi^2(6) = 36.33$, $p < .001$). Model comparisons testing specifically for involvement in an interaction were also significant for polarity



(a) All sentences



(b) True sentences only



(c) False sentences only

Figure 3.3: Mouse-tracking Experiment 1 response accuracy violin plots: the distributions of participant accuracy rates for trials in each condition. Black circles represent means; horizontal lines represent quartiles.

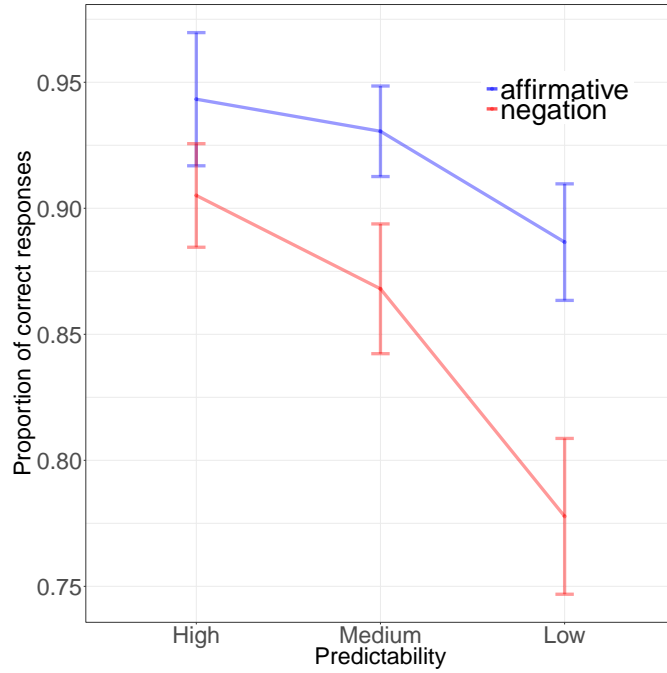


Figure 3.4: Mouse-tracking Experiment

($\chi^2(5) = 11.40$, $p = .044$), predictability ($\chi^2(6) = 16.18$, $p = .013$), and truth value ($\chi^2(5) = 18.60$, $p = .002$). Finally, the model comparison testing for the full three-way interaction indicated that this was present ($\chi^2(2) = 6.19$, $p = .045$).

Coefficients estimated for simple effects in the full model are listed in Table 3.2. In terms of simple interactions, there was a significant interaction between the effects of negative polarity and decreasing from medium to low predictability for false sentences ($\beta = 1.21$, $z = 3.03$, $p = .002$, 95% CI = [0.43, 2.00]), and between the effects of falsity and decreasing from medium to low predictability for negative sentences ($\beta = -0.67$, $z = -2.49$, $p = .013$, 95% CI = [-1.19, -0.143]). No other simple interactions were significant.

Variable	β	z	p	95% CI	
				lower	upper
Falsity					
High pred.					
Aff.	-0.99*	-3.05	.002	-1.62	-0.35
Neg.	-0.22	-0.95	.342	-0.69	0.24
Med. pred.					
Aff.	-0.22	-0.82	.412	-0.75	0.31
Neg.	0.00	0.00	.999	-0.40	0.40
Low pred.					
Aff.	-0.69*	-3.05	.002	-1.13	-0.25
Neg.	-0.67*	-3.87	.001	-1.00	-0.33
Med. pred.					
True					
Aff.	-0.70*	-2.08	.038	-1.36	-0.04
Neg.	-0.28	-1.30	.194	-0.70	0.14
False					
Aff.	0.07	0.26	.795	-0.43	0.57
Neg.	-0.50*	-2.23	.026	-0.95	-0.06
Low pred.					
True					
Aff.	-0.29	-1.08	.282	-0.81	0.24
Neg.	-0.31	-1.57	.117	-0.69	0.08
False					
Aff.	-0.75*	-3.29	.001	-1.20	-0.30
Neg.	-0.97*	-5.32	< .001	-1.33	-0.61
Negation					
True					
High pred.	-1.29*	-3.95	< .001	-1.93	-0.65
Med. pred.	-0.87*	-3.33	.001	-1.38	-0.36
Low pred.	-0.89*	-3.75	< .001	-1.35	-0.42
False					
High pred.	-0.08	-0.29	.771	-0.59	0.44
Med. pred.	-0.65*	-2.61	.009	-1.13	-0.16
Low pred.	-0.86*	-4.44	< .001	-1.24	-0.48

Table 3.2: Mouse-tracking Experiment 1: simple effects on response accuracy of each factor at each level of the other factors. For example, the first row provides the effect of falsity on the proportion of correct responses in a high predictability, affirmative condition. The reference levels of each factor are true (truth value), high (for medium predictability), medium (for low predictability), and affirmative (polarity).

3.3.3 Response initiation time

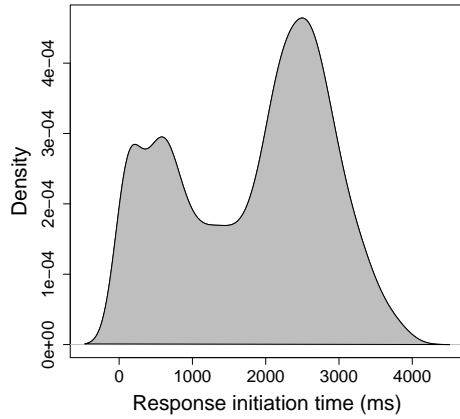
Responses were initiated relatively slowly in this experiment, with participants often taking longer than 2,000 ms to begin moving the mouse. However, as seen in Figure 3.5, there was a clear bimodal distribution of initiation times, with many responses also being initiated much more rapidly. This was confirmed by a significant Hartigan's dip statistic ($D = .06, p < .001$). The largest peak occurred at 2,500 ms. The secondary peak contained two sub-peaks (Figure 3.5a), located at 583 and 573 ms.

Figure 3.6 summarises the differences in initiation time across conditions. As described above, this variable was modelled across all conditions, with a full set of fixed factors and a random effect of polarity included by participant (models involving a more complex random effects structure failed to converge). Model comparisons testing for the presence of main effects and/or interactions were significant for polarity ($\chi^2(6) = 29.82, p < .001$) and for predictability ($\chi^2(8) = 51.42, p < .001$), but not for truth value ($\chi^2(6) = 4.27, p = .640$).

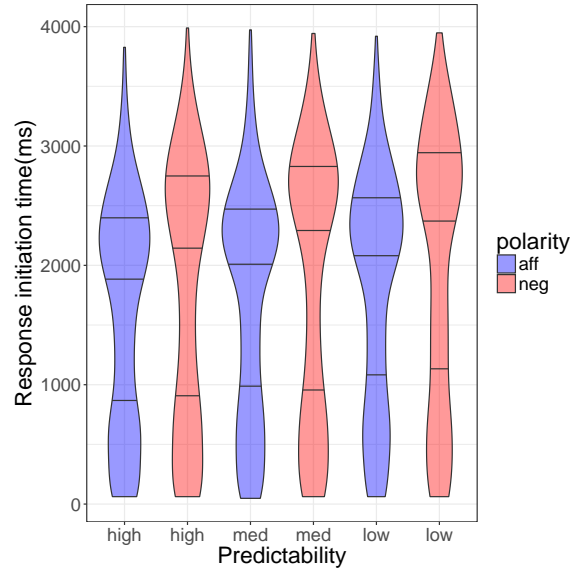
As this pattern indicated the absence of any main effect or interaction involving truth value (including a three-way interaction), the analysis proceeded with a two-way model including only polarity, predictability, and their interaction as fixed factors. This reduced model was compared to a new set of nested models to explore the main effects and interaction of these factors in more detail.

The two-way model was a significantly better fit to the data than a model including only a fixed effect of predictability ($\chi^2(3) = 27.14, p < .001$), confirming a significant main effect of polarity or its involvement in an interaction, and a model including only a fixed effect of polarity ($\chi^2(4) = 47.39, p < .001$), confirming a significant main effect of predictability or its involvement in an interaction. However, the two-way model did not represent a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(2) = 1.55, p = .462$), suggesting a lack of a significant polarity \times predictability interaction.

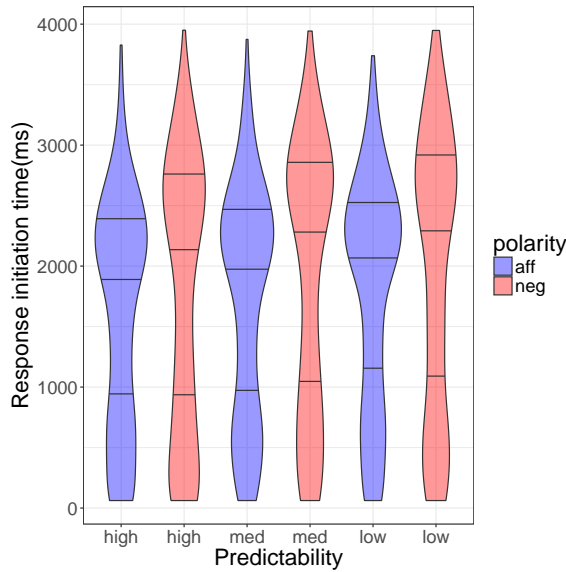
Simple effects and interactions were investigated by estimating coefficients for the full (three-way) model. The former are listed in Table 3.3; there were no significant simple interactions.



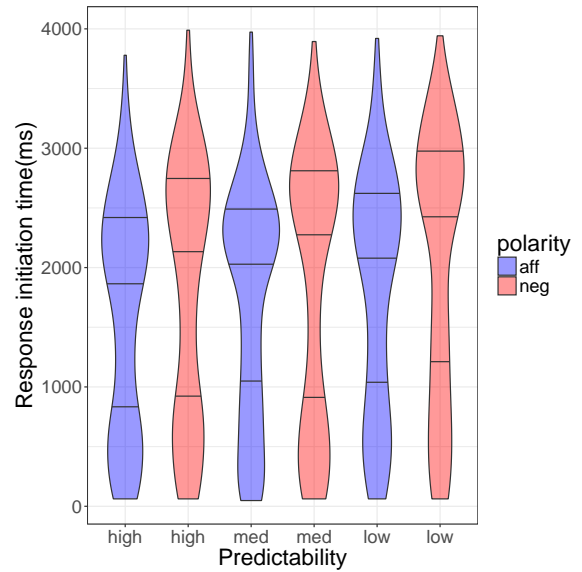
(a) Distribution of response initiation times



(b) All initiation times by condition



(c) True sentences



(d) False sentences

Figure 3.5: Mousetracking Experiment 1 initiation time: distributions of response initiation times, (a) using kernel density estimation, and (b), (c), and (d) across conditions, measured from release of the cursor. Horizontal lines represent quartiles.

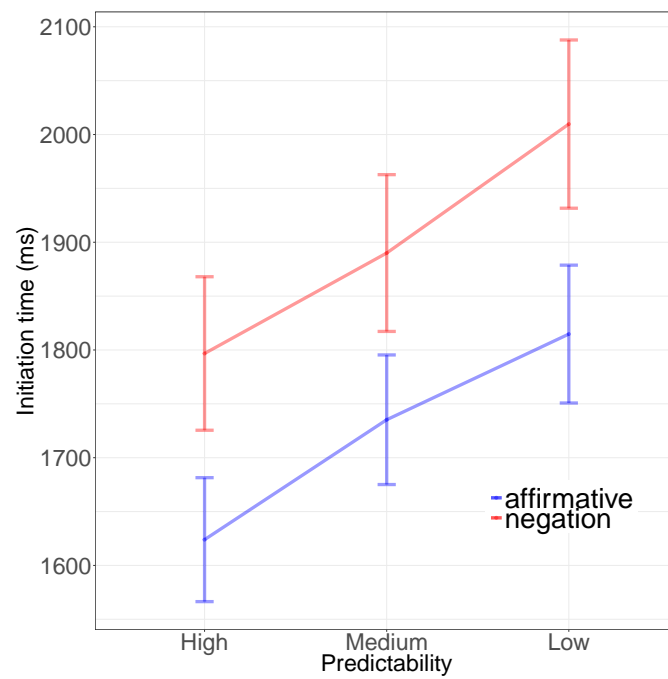


Figure 3.6: Mouse-tracking Experiment 1: mean response initiation time across each level of polarity and predictability, collapsing over truth value. Error bars represent 95% confidence intervals.

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Falsity						
High pred.						
Aff.	-22.76	-0.38	4860.60	.707	-141.40	95.88
Neg.	-52.54	-0.75	4589.37	.395	-173.62	68.54
Med. pred.						
Aff.	3.62	0.06	4589.37	.953	-115.79	123.03
Neg.	-73.61	-1.17	4860.65	.243	-197.29	50.07
Low pred.						
Aff.	-0.86	-0.01	3504.03	.989	-123.29	121.56
Neg.	95.29	1.43	4318.23	.154	-35.65	226.23
Med. pred.						
True						
Aff.	112.30	1.86	4860.59	.062	-5.73	230.33
Neg.	153.19*	2.45	4317.96	.014	30.51	275.86
False						
Aff.	138.68*	2.26	4046.65	.024	18.67	258.68
Neg.	27.04	0.43	4589.43	.664	-95.05	149.12
Low pred.						
True						
Aff.	69.73	1.14	4589.37	.253	-49.83	189.29
Neg.	64.99	1.02	4860.61	.380	-60.03	190.02
False						
Aff.	65.25	1.05	4589.28	.296	-57.02	187.52
Neg.	223.89*	3.54	5132.11	< .001	104.23	363.56
Negation						
True						
High pred.	149.54*	2.37	705.03	.018	25.68	273.40
Med. pred.	190.43*	2.97	735.69	.003	64.75	316.11
Low pred.	185.70*	2.85	740.68	.004	58.10	313.29
False						
High pred.	224.84*	3.53	704.27	< .001	100.13	349.55
Med. pred.	113.20	1.76	743.83	.079	-12.92	239.32
Low pred.	281.85*	4.12	923.56	< .001	147.64	416.06

Table 3.3: Mouse-tracking Experiment 1: effects on initiation time of each factor at each level of the other factors. For example, the first row provides the effect of falsity on the time taken to initiate a response in a high predictability, affirmative condition. The reference levels of each factor are true (truth value), high (for medium predictability), medium (for low predictability), and affirmative (polarity).

3.3.4 Response completion time

Commensurately with the slow initiation times, response completion times were also generally rather slow. After trimming as described above, the distribution still exhibited a positive skew (Figure 3.7) around a peak located at 3,347 ms. However, although Hartigan's dip statistic suggested significant non-unimodality ($D = .001$, $p = .001$), no clear second peak was evident in this case.

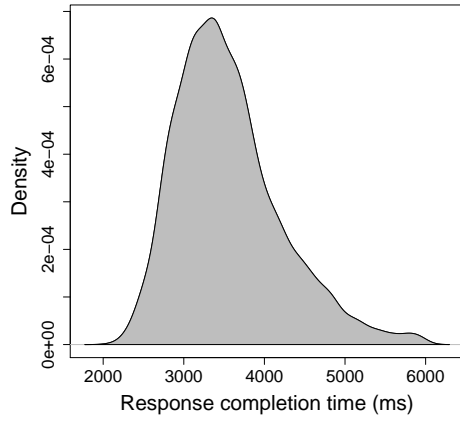
Figure 3.8 summarises the differences in initiation time across conditions. As described above, this variable was modelled across all conditions, with a full set of fixed factors and a random effect of polarity included by participant (models involving a more complex random effects structure failed to converge). Model comparisons testing for the presence of main effects or involvements in an interaction were significant for polarity ($\chi^2(6) = 102.42$, $p < .001$), predictability ($\chi^2(8) = 118.16$, $p < .001$), and truth value ($\chi^2(6) = 85.68$, $p < .001$). Model comparisons indicating involvement in an interaction specifically were significant for polarity ($\chi^2(5) = 32.37$, $p < .001$), and for truth value ($\chi^2(5) = 27.18$, $p < .001$), but not for predictability ($\chi^2(6) = 6.02$, $p = .421$).

As this pattern indicated the absence of any main effect or interaction involving predictability (including a three-way interaction), the analysis proceeded with a two-way model including only polarity, truth value, and their interaction as fixed factors. This reduced model was compared to a new set of nested models to explore the main effects and interaction of these factors in more detail.

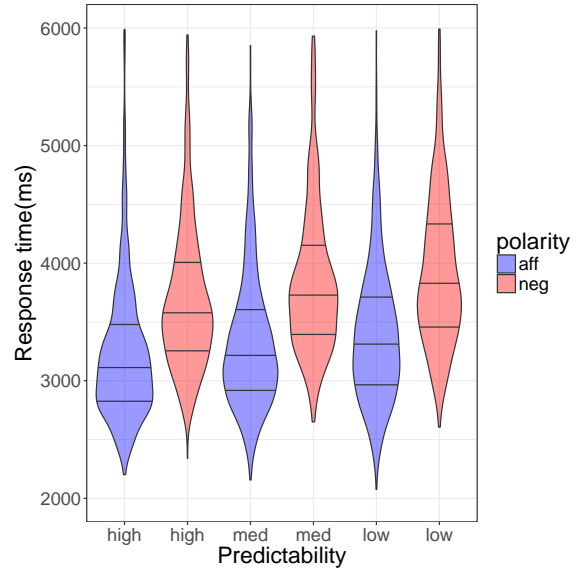
The two-way model was a significantly better fit to the data than a model including only a fixed effect of truth value ($\chi^2(2) = 96.85$, $p < .001$), confirming a significant main effect of polarity or its involvement in an interaction, and a model including only a fixed effect of polarity ($\chi^2(2) = 81.33$, $p < .001$), confirming a significant main effect of truth value or its involvement in an interaction. It was also a significantly better fit to the data than a model including fixed effects of both these factors, and their interaction ($\chi^2(1) = 27.15$, $p < .001$), suggesting a significant polarity \times truth value interaction.

Although there was no three-way interaction, simple effects and interactions were investigated by examining coefficients for the full (three-way) model.

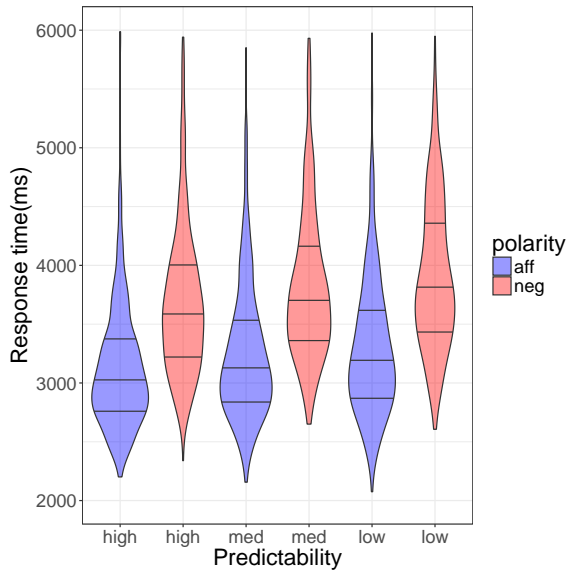
Simple effects on response time are given in Table 3.4. The only significant simple interactions were between negative polarity and falsity at high ($\beta = -204.34$, $t(4589.57) = -3.89$, $p < .001$, 95% CI = $[-307.24, -101.44]$), medium ($\beta = -129.16$, $t(4575.92) = -2.42$, $p = .016$, 95% CI = $[-233.76, -24.55]$), and low ($\beta = -138.17$, $t(4374.21) = -2.48$, $p = .013$, 95% CI = $[-247.24, -29.09]$) predictability.



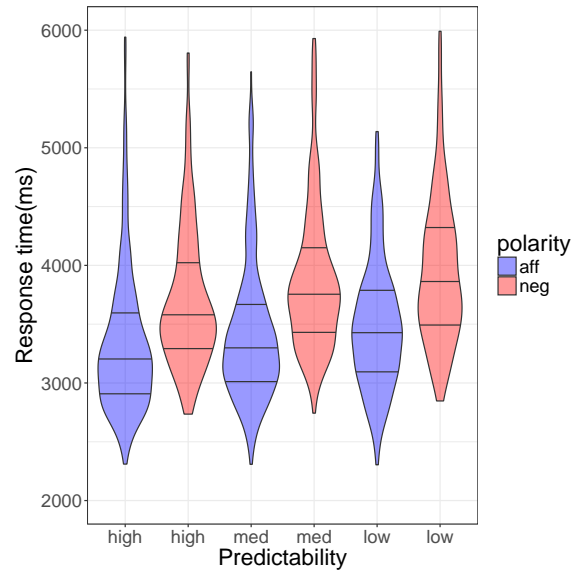
(a) Distribution of response completion times



(b) All completion times by condition



(c) True sentences



(d) False sentences

Figure 3.7: Mouse-tracking Experiment 1: distributions of response completion times, (a) using kernel density estimation, and (b), (c), and (d) for trials in each condition, measured from release of the cursor. Horizontal lines represent quartiles.

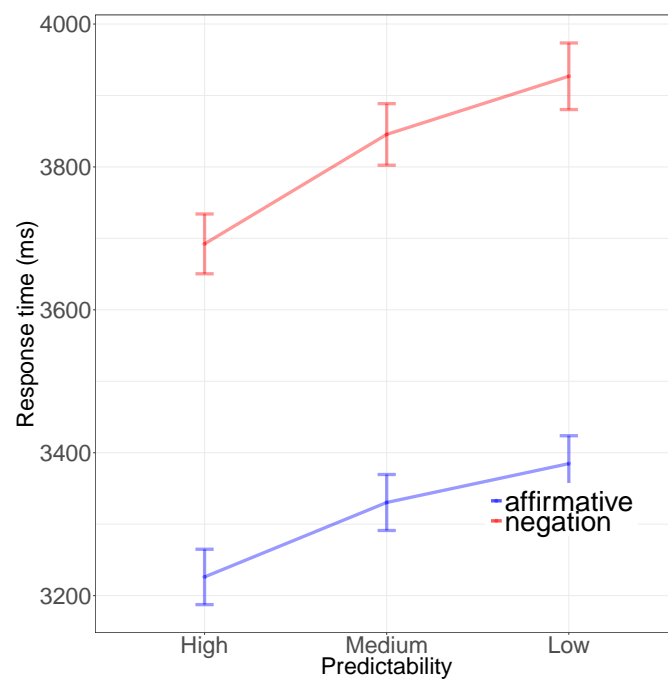


Figure 3.8: Mouse-tracking Experiment 1: mean response completion time across each level of polarity and predictability, collapsing over truth value. Error bars represent 95% confidence intervals.

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Falsity						
High pred.						
Aff.	219.58*	5.98	4591.60	< .001	146.55	291.60
Neg.	15.24	0.41	4485.96	.684	-58.25	88.72
Med. pred.						
Aff.	180.65*	4.87	4573.96	< .001	108.00	253.31
Neg.	51.50	1.34	4576.14	.180	-99.16	114.20
Low pred.						
Aff.	184.94*	4.87	4373.77	< .001	110.45	259.44
Neg.	46.77	1.15	4474.62	.250	-32.90	126.45]
Med. pred.						
True						
Aff.	126.01*	3.36	83.58	.001	52.48	199.54
Neg.	133.53*	3.29	85.49	.001	54.09	212.97
False						
Aff.	86.93*	2.33	4372.94	.020	13.91	159.95
Neg.	169.32*	4.47	4474.84	< .001	95.03	243.61
Low pred.						
True						
Aff.	50.63	1.36	4573.96	.173	-22.11	123.38
Neg.	94.99*	2.45	4574.73	.014	18.92	171.07
False						
Aff.	54.92	1.45	4775.09	.148	-19.47	129.32
Neg.	90.27*	2.24	4476.14	.025	11.36	169.18
Negation						
True						
High pred.	582.51*	13.49	66.50	< .001	497.88	667.14
Med. pred.	589.75*	13.78	166.24	< .001	505.86	673.65
Low pred.	634.11*	14.63	174.49	< .001	549.15	719.07
False						
High pred.	378.21*	8.89	162.02	< .001	294.84	461.57
Med. pred.	460.59*	10.73	168.06	< .001	476.43	544.75
Low pred.	495.94*	10.96	205.36	< .001	407.25	584.63

Table 3.4: Mouse-tracking Experiment 1: simple effects on response completion time of each factor at each level of the other factors. For example, the first row provides the effect of falsity on the time taken to respond in a high predictability, affirmative condition. The reference levels of each factor are true (truth value), high (for medium predictability), medium (for low predictability), and affirmative (polarity).

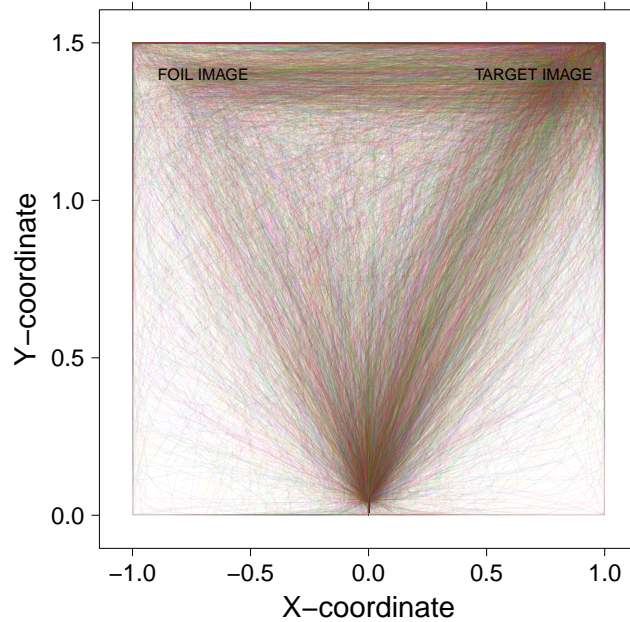


Figure 3.9: All response trajectories in mouse-tracking Experiment 1: a representation of all trials included in the analysis; each line represents the trajectory followed during a single trial, from the starting point in the centre at the bottom of the screen to the finishing point at the target image.

3.3.5 Trajectory shapes

As suggested above by the bimodal distribution of response initiation times, the overall trajectory shapes were not uniform. On many trials, participants moved somewhat directly from the starting point to the target response, taking a relatively straight line. On others, the shape of the trajectory exhibited a degree of “attraction” towards the foil response. This could involve anything from a slightly enhanced curve in the trajectory to a full movement of the cursor all the way to the foil response button and then laterally across the screen to finish at the target.

Figure 3.9 depicts all trials across the experiment in order to demonstrate these types of trajectory. The high density of trials falling in a straight line between the starting and finishing points and in a straight line between the foil and target can be observed. This data is also represented in Figure 3.10, which illustrates how frequently every location on the screen was visited during the course of a trial, across the whole experiment.

Figure 3.11 illustrates the distributions of different trajectory shapes, as sum-

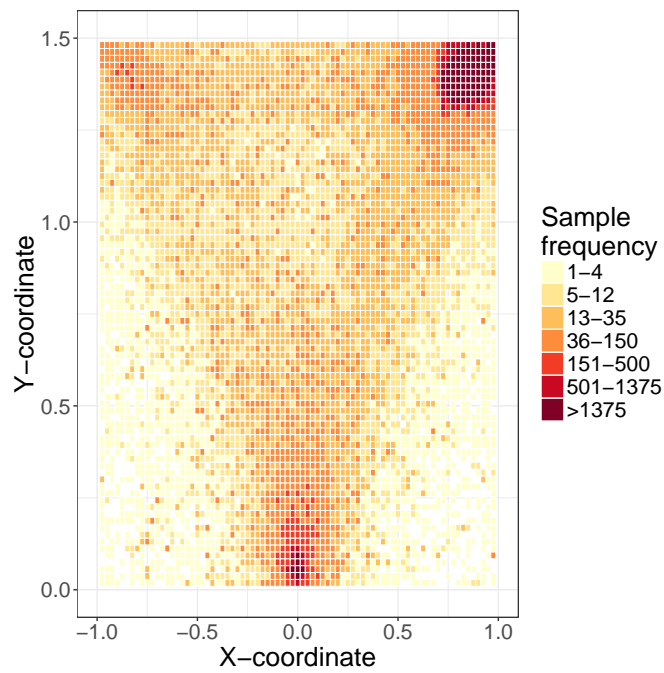


Figure 3.10: Mouse-tracking Experiment 1 heat map of cursor locations: a representation of how frequently, across all trials, any participant's cursor was sampled at each location, using 6,400 (80×80) location bins. Darker colours indicate more frequently visited locations. As in Figure 3.9, the starting point is in the centre at the bottom of the screen, the target response button in the top right corner, and the foil response button in the top left corner. As all included trials were ultimately answered correctly, the extremely dark patches at the starting point and the target response are expected.

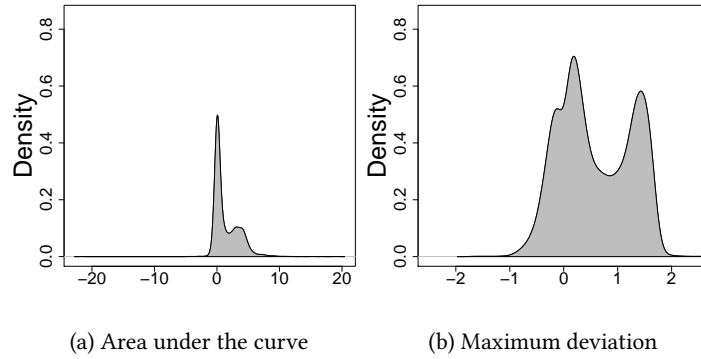


Figure 3.11: Mouse-tracking Experiment 1 area under the curve and maximum deviation: distributions of measures characterising the deviance of individual trials, across all conditions, using kernel density estimation. X axes represent arbitrary units in each case, measuring (a) total area between the trajectory and a straight line between its starting and finishing points, and (b) the distance between the trajectory and the same ideal straight line, at its furthest point.

marised by the AUC and MD. Two clear peaks are evident for both measures: a large one encompassing the majority of trials, with a very small AUC and MD (indicating a relatively direct path from the starting point to the target); and a smaller one representing much more deviant paths. Thus, a significant minority of trials showed evidence of attraction to the foil response. These distributions also appeared to vary across conditions (Figure 3.12).

These characteristics of the data were taken as further evidence that response trajectories tended to fall broadly into two clusters: one representing almost a straight line with an AUC and MD close to zero, and one representing significant attraction towards the foil image, with a correspondingly greater AUC and MD.

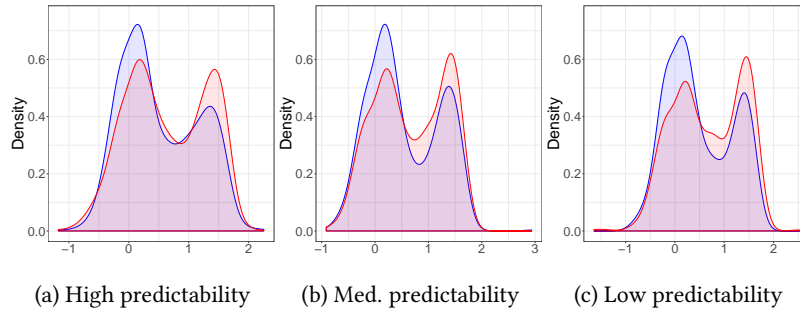


Figure 3.12: Mouse-tracking Experiment 1 distributions of maximum deviation (from a straight line between the starting and finishing points) of trajectories on trials across each condition, using kernel density estimation. Blue plots represent affirmatives; red plots, negations.

3.3.6 Clustering of trajectories

Trajectories were allocated to two clusters, separately for each participant, using a *k*-means algorithm as described above. Figure 3.13 illustrates the two cluster centres that emerged for each individual participant; Figure 3.14 shows the results of these cluster allocations combined for all participants.

Figure 3.15 summarises the pattern of allocations to each cluster for all conditions of polarity and predictability, collapsing across truth value. The proportion of trials falling into each cluster was modelled across all conditions, with a full set of fixed factors and a random effect of polarity included by participant (models involving a more complex random effects structure failed to converge).

Model comparisons testing for the presence of main effects or involvement in an interaction were significant for polarity ($\chi^2(6) = 22.96, p < .001$), predictability ($\chi^2(8) = 77.51, p < .001$), and truth value ($\chi^2(6) = 58.45, p < .001$). Model comparisons indicating involvement in an interaction specifically, however, were not significant for polarity ($\chi^2(5) = 5.10, p = .404$), predictability ($\chi^2(6) = 9.80, p = .133$), or truth value ($\chi^2(5) = 8.18, p = .147$), indicating that only main effects were present for each of the three factors.

Simple effects and simple interactions were investigated by estimating coefficients for the full model. The former are listed in Table 3.5. No simple interactions were present, except the interaction between falsity and decreasing from medium to low predictability for affirmative sentences ($\beta = -0.48, z = -2.08, p = .038, 95\% \text{ CI} = [-0.94, -0.03]$).

A concern for the interpretation of this data might be that trials falling into the

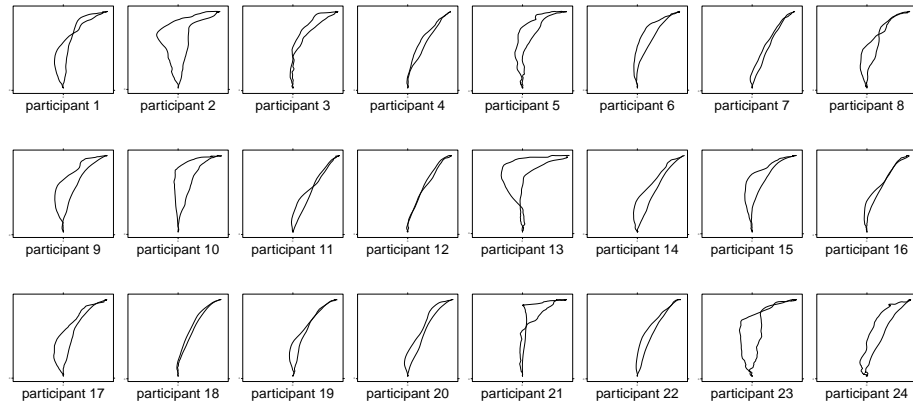


Figure 3.13: Mouse-tracking Experiment 1 individual participant trajectories: the centres of the trajectory clusters identified for each participant in the first stage of the cluster analysis.

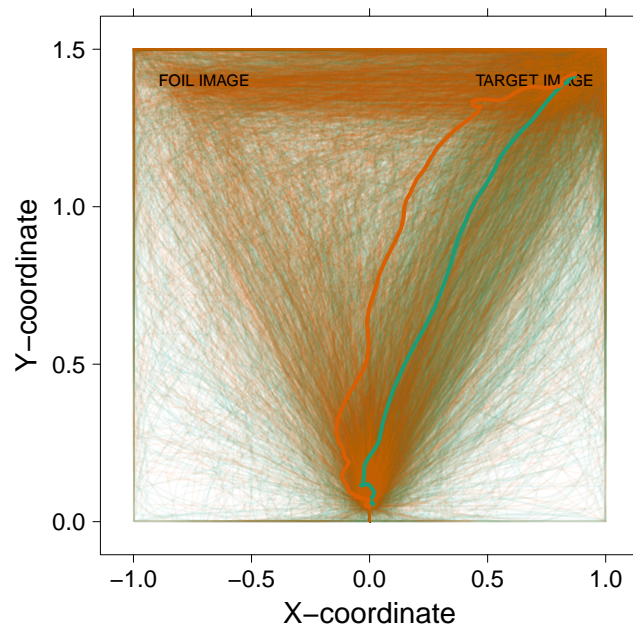


Figure 3.14: Mouse-tracking Experiment 1 trajectory clusters emerging from the cluster analysis. The trajectories of all trials are represented by thin lines (cf Figure 3.9), colour-coded according to the cluster they fall into. Heavy lines represent the corresponding centres of the clusters.

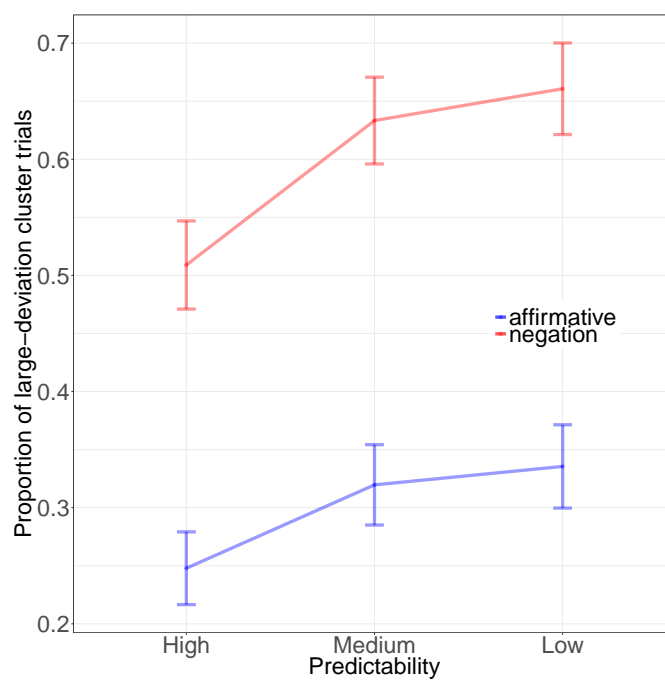


Figure 3.15: Mouse-tracking Experiment 1 effects on cluster trajectory: the mean proportion of trials falling into the more foil-skewed cluster across each condition. Error bars represent 95% confidence intervals.

Variable	β	z	p	95% CI	
				lower	upper
Falsity					
High pred.					
Aff.	-0.35*	-2.01	.044	-0.69	-0.01
Neg.	0.26	1.73	.084	-0.04	0.56
Med. pred.					
Aff.	-0.35*	-2.17	.030	-0.67	-0.03
Neg.	-0.56*	-3.45	.001	-0.88	-0.24
Low pred.					
Aff.	-0.84*	-5.01	< .001	-1.17	-0.51
Neg.	-0.53*	-3.01	.003	-0.87	-0.18
Med. pred.					
True					
Aff.	-0.41*	-2.39	.017	-0.75	-0.07
Neg.	-0.43*	-2.80	.005	-0.73	-0.13
False					
Aff.	-0.42*	-2.52	.012	-0.74	-0.09
Neg.	-0.73*	-4.57	< .001	-1.04	-0.42
Low pred.					
True					
Aff.	0.15	0.90	.367	-0.18	0.49
Neg.	-0.19	-1.20	.230	-0.50	0.12
False					
Aff.	-0.33*	-2.08	.037	-0.64	-0.02
Neg.	-0.16	-0.90	.371	-0.51	0.19
Negation					
True					
High pred.	-1.33*	-4.22	< .001	-1.95	-0.71
Med. pred.	-1.35*	-4.31	< .001	-1.96	-0.73
Low pred.	-1.69*	-5.35	< .001	-2.31	-1.07
False					
High pred.	-1.24*	-3.97	< .001	-1.85	-0.63
Med. pred.	-1.55*	-4.95	< .001	-2.17	-0.94
Low pred.	-1.38*	-4.32	< .001	-2.01	-0.76

Table 3.5: Mouse-tracking Experiment 1: simple effects of each factor on the proportion of trials falling into the more direct cluster at each level of the other factors. For example, the first row provides the effect of falsity on the proportion of direct cluster trials in a high predictability, affirmative condition. The reference levels of each factor are true (truth value), high (for medium predictability), medium (for low predictability), and affirmative (polarity).

more deviant cluster, rather than representing any particular attraction to the foil response, might simply represent trials where the participant happened to initiate their response more quickly by making an initial guess at the correct answer and updating their guess part-way through the trajectory. This would mean that trials falling into the more direct cluster would disproportionately represent those on which the participant hesitated before initiating their response, masking any attraction to the foil.

To rule out this interpretation, initiation times for trials falling into each cluster were compared. In fact, a paired samples t test showed that there was no significant difference in initiation time between clusters ($t(23) = -1.81, p = .08$), and the mean across all conditions was in any case greater for trials falling into the more deviant cluster. This shows that trials falling into the more direct cluster cannot be taken to arise simply as a result of participants pausing or initiating their response more slowly.

3.3.7 Memory test

Before turning to a discussion of the results of the main experiment, the results of the “catch trial” memory test are reported briefly. Participants’ performance on this task was scored by awarding 4 points for each correct item in the correct location; 2 points for each correct item positioned in the wrong location; and 1 point for each incorrect item or X positioned in a location that had contained an item. The total number of points was then divided by the maximum available for that trial (16, 20, or 24 for 4-item, 5-item, and 6-item trials, respectively).

As a percentage of the maximum points available, participants on average scored 76% on 4-item trials, 71% on 5-item trials, and 65% on 6-item trials. A one-way analysis of variance showed that the effect of number of items was significant, $F(2,23) = 18.3, p < .001$.

3.4 Discussion

Two findings are strongly in evidence from the data presented here: the main effects of polarity and predictability. Both negative polarity and decreasing predictability had detrimental effects on all dependent variables, causing participants to respond less accurately, take longer to initiate and complete their correct responses, and follow mouse trajectories more likely to veer initially towards the foil answer. However, the effect of truth value on their responses is less clear: falsity had an

inconsistent, detrimental main effect, with this effect appearing only in the cases of response completion time and trajectory cluster.

In spite of the strength of the principal main effects, the pattern of interactions between variables was difficult to interpret. There was some evidence of a truth value \times polarity interaction (affecting response completion time, with a reduced effect of negation for false sentences), and the suggestion of a three-way interaction among all the independent variables (affecting response accuracy). However, the lack of a clear pattern here makes it difficult to interpret these findings as reliable evidence for any particular conclusion.

Based on previous findings by Nieuwland and Kuperberg (2008) and Dale and Duran (2011), alongside the notion that predictability might function as the “active ingredient” in pragmatic felicity in terms of its interaction with polarity, the hypotheses were that no main effect of polarity or predictability would be observed, but that an interaction between predictability and polarity would show that low predictability hinders incremental processing of negations, generating initially mistaken predictions to a much greater extent than these would appear for affirmatives. None of these hypotheses were strongly supported by the evidence.

First, the main effects of polarity and predictability suggest that neither pragmatic felicity nor predictability can account for all the processing difficulties associated with negation. All sentences presented in this experiment were equally pragmatically felicitous (within the episodic context provided by the visual image and the experimental task). Aside from truth value and polarity, the conditions varied only in predictability of the critical object name and the memory load required to store all objects and their locations in order to make a prediction mid-sentence. Despite this, negation consistently rendered participants slower, less accurate, and more attracted to the foil response option. A possible explanation for the consistent main effect of polarity might be located in the specific details of the context that made these sentences felicitous. Dale and Duran (2011) observed that the main effect of negation only disappeared entirely in the case where they provided the richest versions of their contextual “preambles” (Experiment 3); these sentences were also the most similar to those presented by Nieuwland and Kuperberg (2008) in the licensed condition. In the case of the present experiment, although the visual contexts licensed all the sentences presented, it is possible that their episodic nature interfered with the mechanism facilitating incremental processing of negation in more felicitous contexts. There may be a meaningful distinction, in this case, between the long-term, stable associations in memory between everyday objects

and their properties (for example), and the weak, temporary associations formed between locations and object names for a task such as this one. In this way, the sentences presented here may have been more comparable to those presented by Dale and Duran (2011) in their Experiment 2: felicitous, but only weakly contextually enriched. The pattern of main effects observed matches between these two experiments.

Second, the inconsistent pattern of interactions means that it is difficult to say whether there was an underlying interaction between polarity and predictability in any aspect of processing, as hypothesised. If such an interaction was present, it was very weak (although in the same direction as the hypothesis: i.e., with a reduced impact of negation in higher-predictability sentences), and perhaps masked by two limitations of the experimental design and data analysis: first, the confounding effects of truth value, and second, the weakness of the cluster analysis.

Regarding the confounding effect of truth value, the interaction of truth value judgement and polarity in sentence processing — specifically, that it is easier for participants to judge a negative than an affirmative sentence false — has been consistently reported over a long period (e.g., Gough, 1965; Wason, 1959, 1961). This interaction appears to have been in operation in the present experiment for response completion time and as part of a three-way interaction for response accuracy. This effect is problematic, because in the current experimental design, the manipulation of truth value was not as an independent variable of theoretical interest, but purely a mechanism for producing a task that would tap into participants' ability to process the sentences incrementally and update their predictions for upcoming material. Ideally, the analysis would simply collapse across truth value, ignoring its effects; this would also cancel out the confound between truth value and position of the target and foil response options (TRUE was always presented on the left side of the screen and FALSE on the right), with influences perhaps caused by handedness or reading order forming a component of any main effects of truth value. However, because of the interactions involving truth value, this approach to analysis was not possible, and the effects of the independent variables of interest cannot be fully disentangled from the effects of truth value. Furthermore, inserting the metalinguistic task of truth value judgement into the naturalistic task of sentence processing may have different effects on the responses to true and false sentences, with the latter being easier to verify (see, e.g., Tian & Breheny, 2015, for discussion).

Regarding the cluster analysis, Figures 3.13 and 3.14 illustrate the relative weakness of the trajectory clustering in this experiment. For most of the individual participants, and in the combined analysis, the two cluster centres were not very convincingly distinct. Based on the bimodal distributions of area under the curve and maximum deviation (Figure 3.12) and a visual evaluation of the clustering patterns across all trials (Figures 3.9 and 3.10), the cluster analysis would have been expected to produce two clearly distinguishable clusters, one with a much more pronounced attraction towards the foil response; this was the case for only a small number of participants. This outcome can probably be attributed to two main factors: first, the relatively slow overall response times, meaning that not all stages of initial processing may have been captured by the trajectories; and second, an inadequate number of datapoints. Both these factors imply that the cluster analysis is likely to have underestimated the size of any relevant effects; more specifically, a polarity \times predictability interaction, as hypothesised, could have been present in the underlying processes but have failed to achieve significance under this analysis.

As a result of the above limitations and the overall lack of clarity provided by the results of this experiment, an alternative approach is required in order to better understand the combined effects of polarity and predictability in this type of paradigm. The desiderata for an improved approach are as follows: to avoid the confounding effects of truth value judgement; to elicit faster responses (in terms of both initiation and completion); and to collect more datapoints per participant. The latter two changes would improve the power of the experimental design to capture and detect underlying effects on the shape of the response trajectories. The first requires a more fundamental change to the experimental design; however, this experiment has shown that avoiding truth value judgement is essential for this design.

In addition to the confounding problems discussed above, using truth value judgement as the response task introduces several additional problems. First, there is evidence that sentence processing carried out under a “truth evaluation mindset” may differ in important ways from sentence processing in the absence of this mindset (Wiswede et al., 2013), constraining the generalisability of any findings observed using a truth value judgement task. Second, the dependent variable in the mouse-tracking paradigm does not rely upon a measure that is expected to differ intrinsically according to whether the sentence is initially judged true or false (as in the amplitude of the N400 in an EEG paradigm), but rather upon the behavioural

output of an explicit truth value judgement. Not only is this fact responsible for some aspects of the confounding, it also means that the dependent variable in this experiment relies on a relatively slow, difficult task that requires conscious processing, impairing its ability to access the automatic, rapid cognitive processes that are of greater interest. Finally, the use of truth value judgement as a task may also create an environment in which participants are unusually disincentivised to engage in the very kind of processing that is of interest: incremental prediction-making. Because half of all sentences presented in the experimental context are false, generating predictions on the assumption that a sentence will be true is an unusually poor strategy for generating accurate predictions in this context. Even if comprehenders do generally employ this strategy, its effects might be attenuated (whether consciously or unconsciously) in such a context.

In sum, truth value judgement acts as an additional layer of complexity and set of processing requirements, providing only indirect, noisy, and impaired access to the cognitive content that is truly of interest: participants' predictions for upcoming material in the sentence. In the next chapter, mouse-tracking Experiments 2 and 3 adopt a new methodological approach that avoids the need for truth value judgements, in order to address these limitations and obtain clearer evidence regarding the interaction between polarity and predictability.

Chapter 4

Mousetracking Experiments 2 and 3: Sentence Completion

4.1 Introduction

The findings of mouse-tracking Experiment 1 represented an ambiguous set of results. As discussed in Chapter 3, the interaction of the truth-value judgement task with the independent variables of interest, combined with a perhaps slightly underpowered dataset for the cluster analysis of trajectories, made it difficult to draw firm conclusions. In Experiments 2 and 3, the aim was to improve on the design and methodological approach to obtain unambiguous evidence on the same research question.

Experiments 2 and 3 were therefore based on the same underlying idea as Experiment 1, and adopted broadly the same methodology. However, the truth-value judgement task was, crucially, replaced with a sentence completion task. In this paradigm, rather than hearing complete sentences and judging their truth or falsity, participants were presented with incomplete sentence fragments; their task was to select the picture of the object that would truthfully complete the sentence fragment, given the visual context of the trial.

This task offered several advantages over the truth value judgement task used in Experiment 1. First, it was expected to be a quicker and easier task for participants to perform, freeing up cognitive resources previously expended on the task itself — perhaps for use in more accurate or more difficult prediction-making. Since explicit truth-value judgements are not part of ordinary sentence processing, this should also more closely approximate non-laboratory conditions. Second,

and relatedly, this task should more directly tap the actual cognitive process of interest, namely prediction. Relatively few assumptions are required to expect that participants might exhibit initial attraction towards their predicted completion for the sentence fragment, in comparison to the additional layer of processing strategy required to make this assumption in relation to truth value judgement. The task should also incentivise participants to make predictions to a greater extent than in Experiment 1, which contained as many false as true sentences. Third, because the response options in the picture choice task change on every trial, it is possible to counterbalance their location on the screen (left or right), thus avoiding potential confounds associated with TRUE responses always involving a leftward movement, and FALSE responses always a rightward movement. Finally, on a relatively minor note affecting only a few experimental trials, the use of images avoids problems with the naming of items. (Despite the careful selection of images, it was apparent from the memory test responses in Experiment 1 that participants' names for object sometimes did not match those used in the sentences: for instance, in the absence of a clear scale for size, a *bracelet* was frequently mistaken for a *necklace*.) Because the sentence fragments never contain the names of any items, participants are free to make predictions on the basis of their own labels for them, without affecting the data.

It was anticipated that the sentence completion task would yield quicker response initiation and completion times across the board compared to the truth value judgement task. However, it was important to be sure of collecting data as early in participants' cognitive processing stages as possible. Therefore, Experiment 3 was simply a speeded version of Experiment 2, with added auditory cues to encourage participants to begin and complete their responses as quickly as possible.

4.2 Methods

4.2.1 Materials

The visual materials were identical for Experiments 2 and 3. A set of 216 images of everyday objects, selected for their ease and consistency of recognition and naming, were used in creating the stimuli. Each stimulus consisted of a 3×2 grid containing 4, 5 or 6 of these object images, paired with a sentence fragment that was presented auditorily. The sentence fragment described a target location so that the sentence could be completed by the participant having observed the associated

image.

In total, 216 sets of 8 stimuli were prepared such that each participant would see 1 stimulus from each set. Within a set, 4 stimuli fell into critical conditions (2 with affirmative sentence fragments and 2 with negative) and the other 4 were used in “constant-predictability” trials; their predictability was always low (because the row referred to was always the one containing three objects). These were included for counterbalancing purposes (so that participants would be unable to predict which object would be the target answer by its position alone). In each affirmative/negative pair of conditions, the sentence fragment and the set of objects in the image were kept the same, but the locations of two critical objects in the image were exchanged in order to counterbalance the identity of the target.

Each participant in each experiment thus completed 108 critical trials (18 in each of 6 conditions: affirmative or negative sentence fragment crossed with group size of 4, 5 or 6 for high, medium or low predictability of target), 108 constant-predictability counterbalancing trials, and also 108 filler trials, each using a different group of images (although individual objects were repeated, an equal number of times, between trials).

For critical and constant-predictability conditions, all sentence fragments took the form: “The top/bottom row contains/doesn’t contain...”. For filler trials, the sentence fragment referred instead to the *left side* or *right side*. Audio recordings of these fragments (which were also identical for Experiments 2 and 3) were prepared and edited in the same way as for Experiment 1. Table 4.1 summarises how the Experiment 2 and 3 stimuli were constructed.

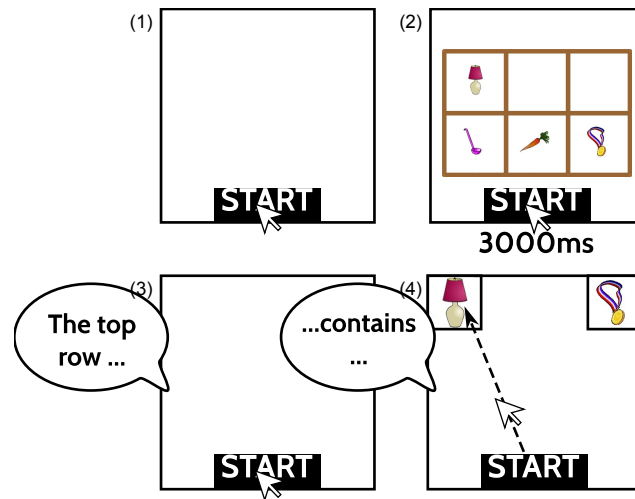
Additionally, a 2,000 ms audio clip was created in Audacity for use in Experiment 3 only. This consisted of an initial beep to signal that the participant should begin their answer followed by a series of beeps, at decreasing intervals and increasing in amplitude, intended to increase the participant’s sense of time pressure. For Experiment 3, this clip was superimposed onto each sentence fragment audio to produce the final auditory stimuli, so that the initial signal beep coincided with the onset of the critical word (*contains* or *doesn’t*). The clip began at a low amplitude so that the critical word was not masked and remained clearly audible.

No catch trials (the grid recall task) were included in Experiments 2 or 3. Based on the slow response times, overall accuracy rates, and informal reports of participants, it was judged that participants may have been over-expending resources on the memory task, to the detriment of their performance in the primary task. Instead, the inclusion of the constant-predictability trials provided another

Predictability	Image	Condition	Sentence fragment	Target	Foil
High		critical affirmative	The top row contains...		
		critical negative	The bottom row doesn't contain...		
		counterbalance affirmative	The bottom row contains...		
		counterbalance negative	The top row doesn't contain...		
		critical affirmative	The top row contains...		
		critical negative	The bottom row doesn't contain...		
		counterbalance affirmative	The bottom row contains...		
		counterbalance negative	The top row doesn't contain...		
Medium		critical affirmative	The bottom row contains...		
		critical negative	The top row doesn't contain...		
		counterbalance affirmative	The top row contains...		
		counterbalance negative	The bottom row doesn't contain...		
		critical affirmative	The bottom row contains...		
		critical negative	The top row doesn't contain...		
		counterbalance affirmative	The top row contains...		
		counterbalance negative	The bottom row doesn't contain...		
Low		critical affirmative	The top row contains...		
		critical negative	The bottom row doesn't contain...		
		counterbalance affirmative	The bottom row contains...		
		counterbalance negative	The top row doesn't contain...		
		critical affirmative	The top row contains...		
		critical negative	The bottom row doesn't contain...		
		counterbalance affirmative	The bottom row contains...		
		counterbalance negative	The top row doesn't contain...		

Table 4.1: Mouse-tracking Experiments 2 and 3 examples stimulus sets. One example set is shown for each level of predictability. Each pair of grid images produced 8 trial conditions (4 critical, 4 for counterbalancing purposes), 1 of which was seen by any given participant. The appearance of empty cells on the top or bottom row, and thus the conditions that particular sentence fragments fell into, was randomly allocated (in equal numbers) to pairs of images.

Figure 4.1: Mouse-tracking Experiments 2 and 3 example trial. Participant clicks the START button; visual stimulus is displayed; sentence fragment is played; at onset of critical word, response buttons appear and participant makes selection using the mouse. In Experiment 3, the escalating beep signal applying pressure to respond also begins at the same time as the onset of the critical word. Stimuli and response buttons are for illustration and not to scale.



avenue by which to assess the effects of memory load (see Discussion).

4.2.2 Procedure

The procedure was identical to Experiment 1, except in the following respects. During each trial, between 690 and 902 ms after the start of the audio, depending on its contents, the onset of the critical word (*contains* or *doesn't*) occurred, and at the same time, the response options appeared as images located in the top-right and top-left corners of the screen. The response signal tone also occurred and the mouse cursor was released at this time. In Experiment 2, participants were required to initiate their response (by beginning a mouse movement) within 5,000 ms of this time; in Experiment 3, they were required to complete their response (by clicking on either the target or the foil image) within 2,000 ms of this time. In either case, if they failed to do so, they received a warning message. After completing their response, the participant received feedback in the form of a green O (correct) or red X (incorrect) displayed for 300 ms. An example trial is shown in Figure 4.1.

Participants completed two practice trials, followed by the 324 experimental trials (critical, constant-predictability, and filler conditions) separated into 9 blocks

of 36 trials and presented in a random order. Mouse coordinates were sampled online every 20 ms during each trial; other dependent variables were measured in the same way as for Experiment 1.

4.2.3 Participants

In both experiments, informed consent was obtained from all participants prior to their participation, and they were compensated for their time with course credit or a small payment. Ethical approval for Experiments 2 and 3 was granted by the University of Bristol's Faculty of Science Research Ethics committee (ethical approval code 35142).

Experiment 2

Twenty-four participants, none of whom had participated in mouse-tracking Experiment 1, were recruited from in and around the University of Bristol community (6 male, aged 18–23 years [$M = 19.7$, $SD = 1.1$]). All were native speakers of English and all but three (USA/France, Cyprus, Switzerland) had grown up primarily in the United Kingdom. Most were monolingual but 11 had second languages of a good, advanced or fluent standard (Cantonese, Portuguese, French, Spanish, Greek, German, Dutch and Turkish represented). Due to experimenter error, age was not recorded for one participant and age, country during childhood, and additional languages were not recorded for a second participant.

Experiment 3

Thirty-three participants, none of whom had participated in mouse-tracking Experiments 1 or 2, were recruited from in and around the University of Bristol community. All were native speakers of English, and all but six (France (2), Macau, Hong Kong, Malaysia and Ireland) had grown up primarily in the United Kingdom. Most were monolingual but 10 had second languages of a good, advanced or fluent standard (French, Polish, Arabic, Cantonese, Mandarin, Italian and Malay represented).

One participant was rejected from the analyses due to anomalously high error rate (36%, compared to a mean error rate among all other participants of 12%). Of the remaining 32 participants, nine were male and the age range was 18–34 years ($M = 23.5$, $SD = 4.5$).

4.3 Results

4.3.1 Data preparation and analysis

Among all participants in Experiment 2 ($N = 24$), the range of error rates across all critical conditions was 0.5% to 51%. In Experiment 3 ($N = 32$), the range was 1% to 26%.

Filler trials and those included for counterbalancing purposes were excluded from the analysis for both experiments. From a total of 2,592 critical trials (108 per participant) in each experiment, those with incorrect responses were discarded (15% of critical trials in Experiment 2, and 12% in Experiment 3). For Experiment 2, trials with response completion times longer than 4,000 ms, and on which the participant took longer than 2,000 ms to initiate their response, were also excluded (a further 6% of the remaining critical trials). For Experiment 3, the bounds were at 2,000 ms and 1,000 ms, respectively, which led to exclusion of a further 8% of the remaining critical trials after the exclusion of incorrectly answered trials. These thresholds were selected on the basis of visual inspection of the distribution of initiation and completion times (see Figures 4.4, 4.6, 4.17, and 4.20 below).

Data from trials in which the target response was located on the left side of the screen were reflected in the same way as for Experiment 1. All other aspects of data preparation and computation of dependent measures for both Experiments 2 and 3 were the same as for Experiment 1.

4.3.2 Statistical modelling of effects

The same approach to modelling of each dependent measure was applied to both Experiments 2 and 3 as for Experiment 1; however, because truth value was not manipulated in the former cases, it was not included as a factor for any analysis. The procedure for model comparison to establish the presence of main effects and interactions was therefore simplified by the presence of only two factors:

1. The full model (including fixed effects of polarity, predictability, and their interaction) was compared to models including only the fixed effect of predictability, and only the fixed effect of polarity. (An example of the formula used for this model, using response time as the dependent variable and a maximal random effects structure, is given in formula 4.1.) If the full model represented a significantly better fit to the data, this was taken as an indication of either a main effect, or the involvement in an interaction, of predictability or polarity, respectively.

2. To test for a polarity \times predictability interaction, the full model was compared to a model including only the fixed effects of each factor, and not their interaction.

$$\begin{aligned} \text{response time} \sim & \text{polarity} * \text{predictability} \\ & + (\text{polarity} * \text{predictability} | \text{participant}) \end{aligned} \quad (4.1)$$

Following these tests, the full model was used to estimate coefficients for the simple effects of each factor at each level of the other factor.

4.3.3 Experiment 2

Response accuracy

Participants were generally able to complete the task accurately across all conditions, although these varied in difficulty (Figure 4.2).

Figure 4.3 summarises the differences in response accuracy across conditions. Accuracy was modelled across all conditions, with a full set of fixed factors and a random intercept by participants (models involving a more complex random effects structure failed to converge). A set of models nested in the full model were compared to explore the main effects. Model comparisons testing for the presence of main effects or involvement in an interaction were significant for both polarity ($\chi^2(3) = 17.6, p < .001$) and predictability ($\chi^2(4) = 32.2, p < .001$). However, the full model did not represent a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(2) = 3.8, p = .15$), indicating a lack of a significant polarity \times predictability interaction.

Although there was no interaction, simple effects were investigated by estimating coefficients for the full model as part of planned comparisons. These are listed in Table 4.2.

Response initiation time

Participants generally started to move the mouse within 500 ms of the cursor being released, and often much more quickly, across all conditions, although there was variation in the shape of the distribution by condition (Figure 4.4). In general, initiation times were distributed bimodally (Figure 4.4a), as confirmed by a significant Hartigan's dip statistic ($D = .046, p < .001$), with peaks at 333 and 63 ms.

Figure 4.5 summarises the variation in initiation time across conditions. This was modelled across all conditions, with a full set of fixed effects and random effects

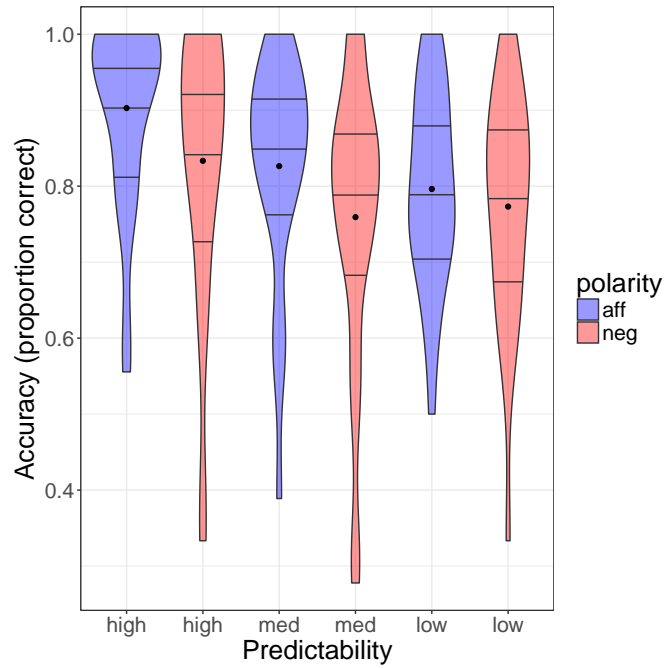


Figure 4.2: Mouse-tracking Experiment 2 response accuracy violin plots: the distributions of participant accuracy rates for trials in each condition. Black circles represent means; horizontal lines represent quartiles. The violin plots illustrate the skewness of the data, to different extents across different conditions.

Variable	β	z	p	95% CI	
				lower	upper
Negation					
High pred.	-0.68*	-3.15	.002	-1.11	-0.26
Med. pred.	-0.47*	-2.59	.01	-0.82	-0.11
Low pred.	-0.16	-0.89	.38	-0.50	0.19
Med. pred.					
Aff.	-0.74*	-3.43	.001	-1.16	-0.32
Neg.	-0.52*	-2.87	.004	-0.88	-0.17
Low pred.					
Aff.	-0.22	-1.21	.23	-0.58	-0.14
Neg.	0.09	0.52	.605	-0.25	0.42

Table 4.2: Mouse-tracking Experiment 2: simple effects of each factor on the proportion of correct responses at each level of the other factors. For example, the first row provides the effect of negative polarity on the proportion of correct responses in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

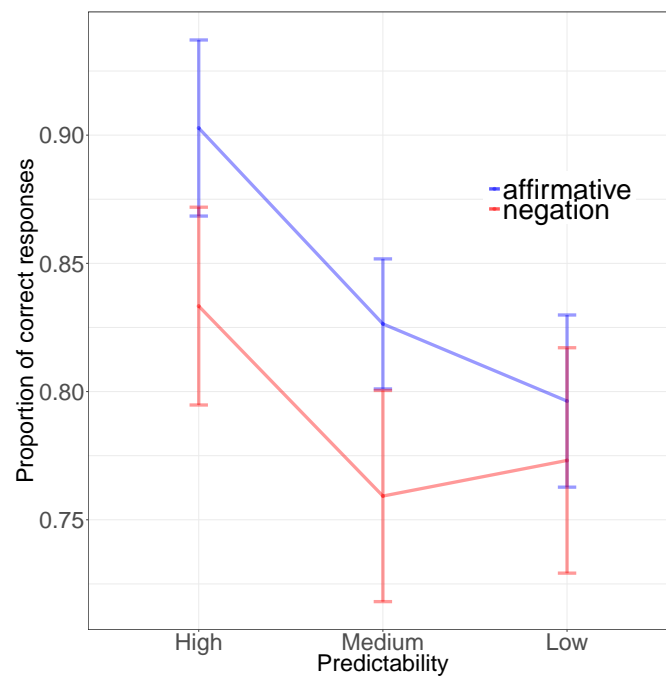


Figure 4.3: Mouse-tracking Experiment 2 response accuracy effects: mean proportion of correct responses across all conditions. Error bars represent 95% confidence intervals.

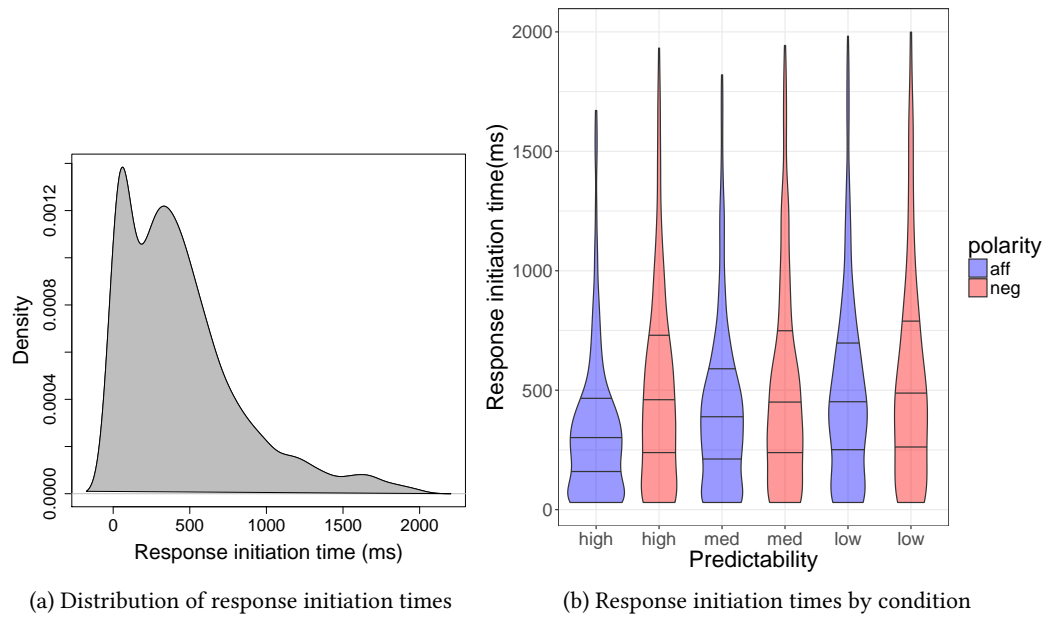


Figure 4.4: Mouse-tracking Experiment 2: distributions of response initiation times, (a) using kernel density estimation, and (b) across conditions, measured from release of the cursor. Horizontal lines represent quartiles.

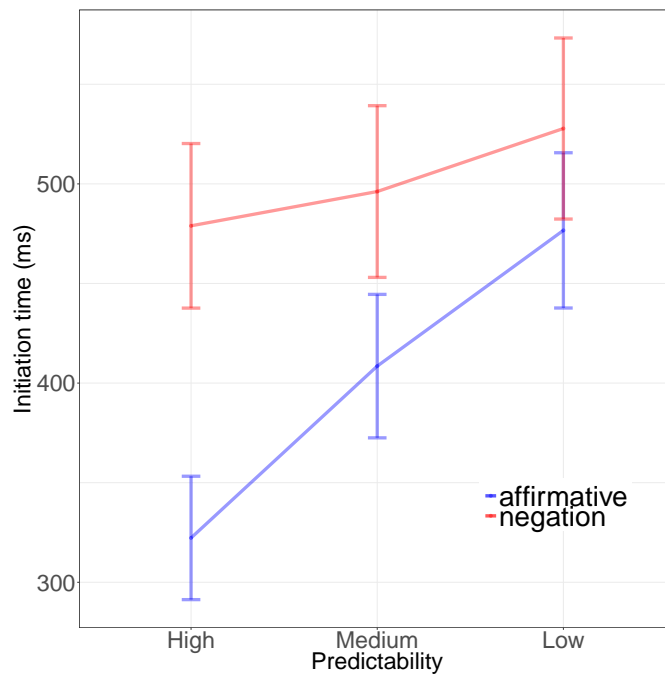


Figure 4.5: Mouse-tracking Experiment 2: mean response initiation time across all conditions. Error bars represent 95% confidence intervals.

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Negation						
High pred.	150.44*	5.35	109.12	< .001	95.38	205.51
Med. pred.	101.82*	3.49	124.25	< .001	44.69	158.95
Low pred.	65.51*	2.24	127.28	.03	8.20	122.82
Med. pred.						
Aff.	82.84*	3.03	101.57	.003	29.33	136.36
Neg.	34.22	1.21	114.75	.229	-21.26	89.69
Low pred.						
Aff.	72.38*	2.76	276.02	.006	20.89	123.87
Neg.	36.07	1.33	296.37	.183	-16.94	89.07

Table 4.3: Mouse-tracking Experiment 2: simple effects of each factor on response initiation time at each level of the other factors. For example, the first row provides the effect of negative polarity on initiation time in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

by participant of both polarity and predictability, but not their interaction. (Models including the interaction in the random effects structure failed to converge.) Model comparisons testing for the presence of main effects or interactions were significant for polarity ($\chi^2(3) = 25.4$, $p < .001$) and predictability ($\chi^2(4) = 24.3$, $p < .001$). However, the full model was not a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(2) = 5.6$, $p = .06$), indicating that the polarity \times predictability interaction was (marginally) not significant.

Although there was no interaction, simple effects were investigated by estimating coefficients for the full model. These are listed in Table 4.3.

Response completion time

The distribution of response completion times (after trimming as described above) was slightly skewed around a peak at 1,118 ms (Figure 4.6); although Hartigan's dip statistic suggested significant non-unimodality ($D = .025$, $p = .03$), no clear second peak was evident in this case (Figure 4.6a).

Figure 4.7 summarises the variation in response completion time across conditions. This was modelled across all conditions, with a full set of fixed effects and the same random effects structure as for initiation time (models including the interaction in the random effects structure again failed to converge). Model

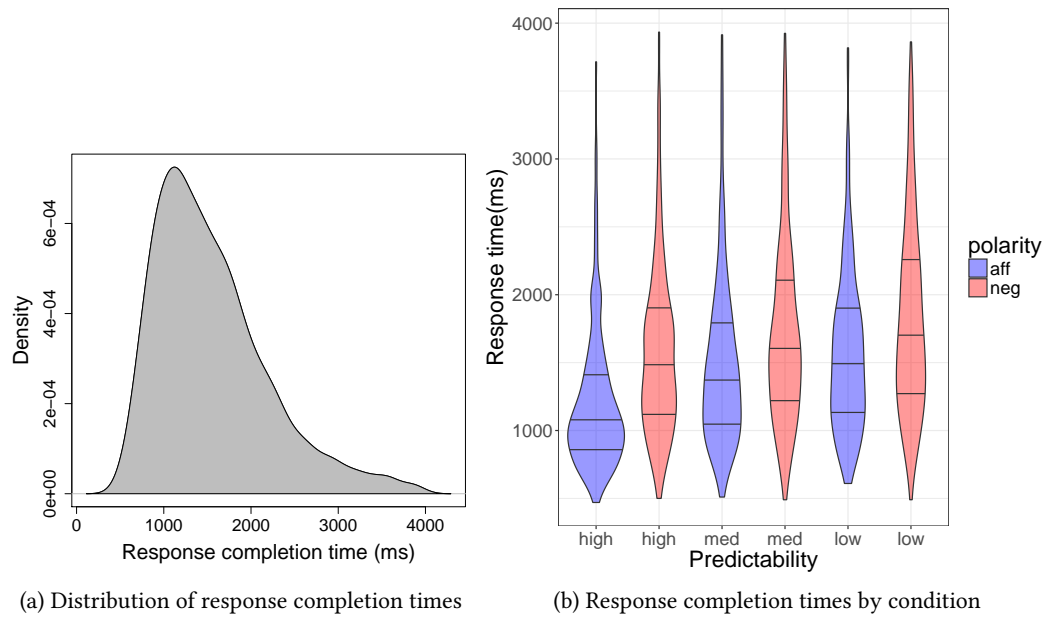


Figure 4.6: Mouse-tracking Experiment 2: distributions of response completion times, (a) using kernel density estimation, and (b) for trials in each condition, measured from release of the cursor. Horizontal lines represent quartiles.

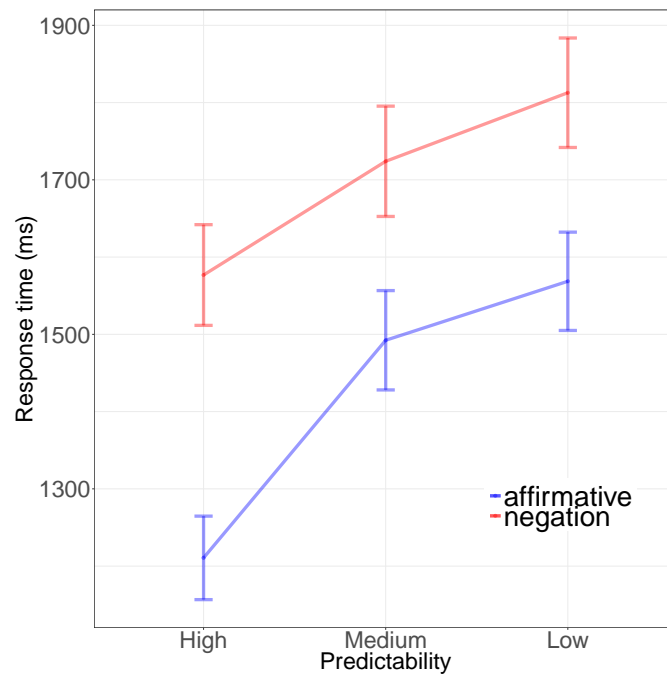


Figure 4.7: Mouse-tracking Experiment 2: mean response time across all conditions. Error bars represent 95% confidence intervals.

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Negation						
High pred.	358.97*	7.38	74.29	< .001	263.69	454.25
Med. pred.	243.46*	4.84	84.20	< .001	144.92	341.99
Low pred.	253.46*	5.02	85.72	< .001	154.28	351.75
Med. pred.						
Aff.	276.05*	5.86	60.16	< .001	183.78	368.33
Neg.	160.54*	3.30	67.37	.002	65.10	255.99
Low pred.						
Aff.	92.35	1.99	62.35	.051	1.30	183.40
Neg.	101.91*	2.14	67.65	.036	8.44	195.37

Table 4.4: Mouse-tracking Experiment 2: simple effects of each factor on response completion time at each level of the other factors. For example, the first row provides the effect of negative polarity on response time in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

comparisons testing for the presence of main effects or interactions were significant for both polarity ($\chi^2(3) = 36.1$, $p < .001$) and predictability ($\chi^2(4) = 40.9$, $p < .001$). However, the full model did not represent a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(1) = 4.7$, $p = .09$), indicating that the interaction was not significant.

Simple effects were investigated by estimating coefficients for the full model. These are listed in Table 4.4.

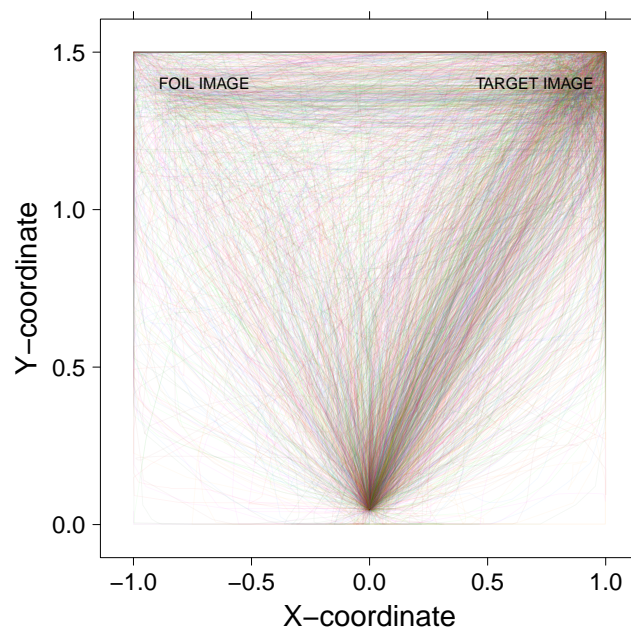


Figure 4.8: Mouse-tracking Experiment 2: a representation of all trials included in the analysis; each line represents the trajectory followed during a single trial, from the starting point in the centre at the bottom of the screen to the finishing point at the target image.

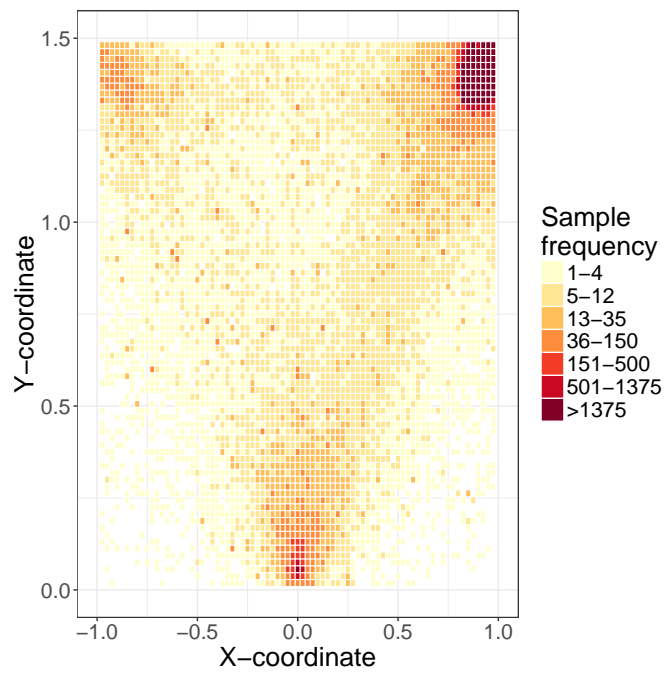


Figure 4.9: Mouse-tracking Experiment 2 heat map of cursor locations: a representation of how frequently, across all trials, any participant's cursor was sampled at each location, using 6,400 (80×80) location bins. Darker colours indicate more frequently visited locations. As in Figure 4.8, the starting point is in the centre at the bottom of the screen, the target image in the top right corner, and the foil image in the top left corner.

Trajectory shapes

The patterns described above, with significant bimodality especially in initiation times, again suggested some level of clustering in the trajectory shapes; see also Figures 4.8 and 4.9, which represent all trial trajectories and all datapoints sampled throughout the experiment, respectively.

This picture is further supported by clear bimodalities in the AUC and MD, computed for each trial as described above (Figures 4.10 and 4.11). The clustering analysis using $k = 2$ therefore remained a valid approach to analysis for this experiment.

Clustering of trajectories

K -means clustering was carried out as described above, yielding two clusters of trajectories. Figure 4.12 illustrates the two cluster centres that emerged for each individual participant; Figure 4.13 shows the results of these cluster allocations combined for all participants.

Figure 4.14 shows the proportion of trials falling into each cluster across conditions. The proportion of direct-cluster trials was modelled across all participants, with a full set of fixed factors and a random effect of polarity included by participant (models with a more complex random effects structure failed to converge).

Model comparisons testing for the presence of main effects or interactions were significant for both polarity ($\chi^2(3) = 30.0, p < .001$), and predictability ($\chi^2(4) = 88.7, p < .001$). The full model was also a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(2) = 10.5, p = .005$), indicating a significant polarity \times predictability interaction.

Simple effects were investigated by estimating coefficients for the full model. These are given in Table 4.5.

As in Experiment 1, a concern for the interpretation of this data might be that trials falling into the more deviant cluster, rather than representing any particular attraction to the foil response, might simply represent trials where the participant happened to initiate their response more quickly by making an initial guess at the correct answer and updating their guess part-way through the trajectory. To rule out this interpretation, initiation times for trials falling into each cluster were compared. In fact, the mean initiation time was greater for trials falling into the more deviant cluster, and a paired samples t test showed that this difference was significant ($t(23) = -6.69, p < .001$).

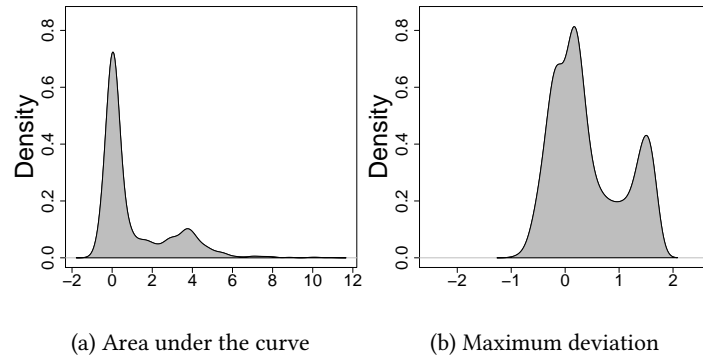


Figure 4.10: Mouse-tracking Experiment 2 area under the curve and maximum deviation: distributions of measures characterising the deviance of individual trials, across all conditions, using kernel density estimation. X axes represent arbitrary units in each case, measuring (a) total area between the trajectory and a straight line between its starting and finishing points, and (b) the distance between the trajectory and the same ideal straight line, at its furthest point.

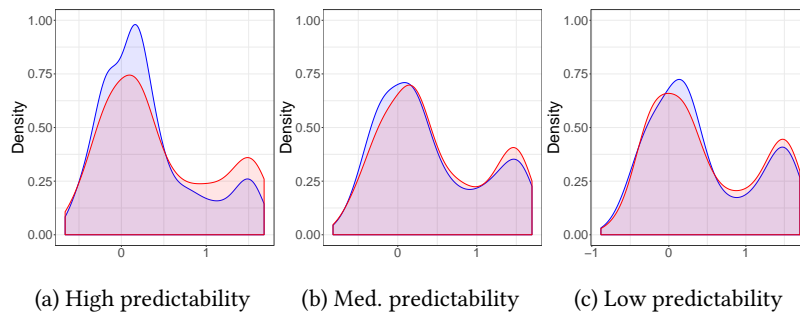


Figure 4.11: Mouse-tracking Experiment 2: distributions of maximum deviation (from a straight line between the starting and finishing points) of trajectories on trials across each condition, using kernel density estimation. Blue plots represent affirmatives; red plots, negations.

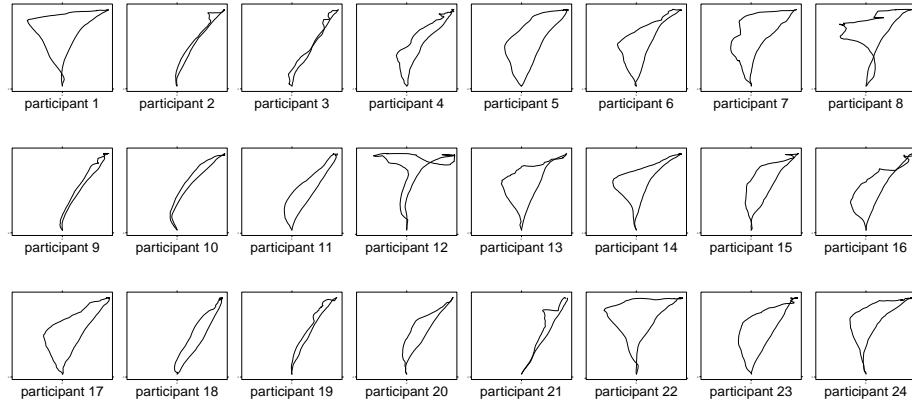


Figure 4.12: Mouse-tracking Experiment 2 individual participant trajectories: the centres of the trajectory clusters identified for each participant in the first stage of the cluster analysis.

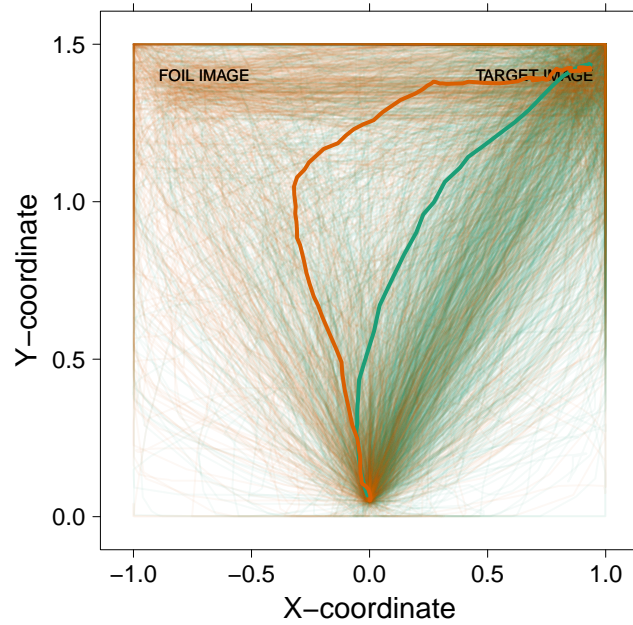


Figure 4.13: Mouse-tracking Experiment 2 trajectory clusters emerging from the cluster analysis. The trajectories of all trials are represented by thin lines (cf Figure 4.8), colour-coded according to the cluster they fall into. Heavy lines represent the corresponding centres of the clusters.

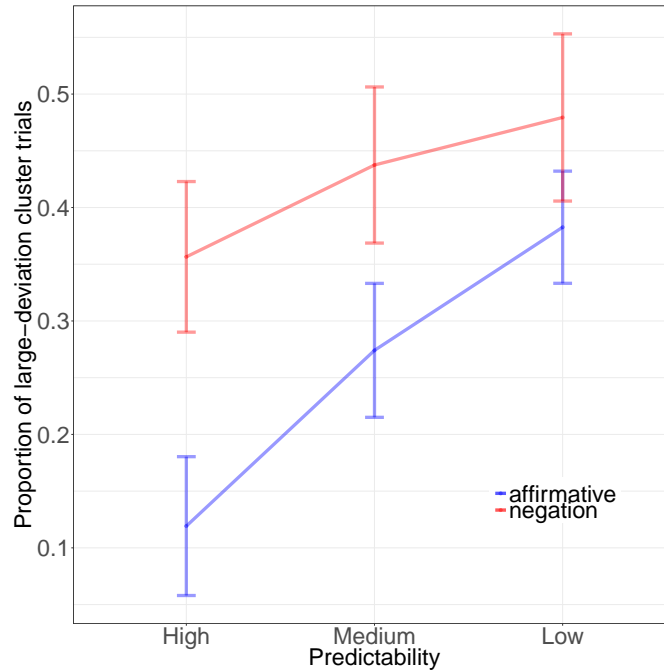


Figure 4.14: Mouse-tracking Experiment 2 effects on cluster trajectory: the mean proportion of trials falling into the more foil-skewed cluster across each condition. Error bars represent 95% confidence intervals.

Variable	β	z	p	95% CI	
				lower	upper
Negation					
High pred.	-1.37*	-6.02	< .001	-1.81	-0.92
Med. pred.	-0.79*	-3.90	< .001	-1.19	-0.39
Low pred.	-0.50*	-2.55	.01	-0.89	-0.12
Med. pred.					
Aff.	-1.07*	-5.22	.001	-1.47	-0.67
Neg.	-0.50*	-2.80	.005	-0.85	-0.15
Low pred.					
Aff.	-0.53*	-3.02	.003	-0.87	-0.19
Neg.	-0.24	-1.34	.180	-0.59	0.11

Table 4.5: Mouse-tracking Experiment 2: simple effects of each factor on the proportion of trials falling into the more direct cluster at each level of the other factors. For example, the first row provides the effect of negative polarity on the proportion of direct-cluster responses in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

4.3.4 Experiment 3

Response accuracy

Based on analysis of their responses, participants were again generally able to complete the task accurately across all conditions, although some were easier than others (Figure 4.15). The mean accuracy rate (88%) was slightly higher than in Experiment 2 (85%), indicating that the increase in time pressure had no detrimental effect on accuracy.

Figure 4.16 summarises the variation in response accuracy across conditions. This was modelled across all conditions with a full set of fixed factors and a random factor of polarity included by participant. (Models involving a more complex random effects structure failed to converge.) Model comparisons testing for the presence of main effects or interactions were significant for both polarity ($\chi^2(3) = 32.5, p < .001$) and predictability ($\chi^2(4) = 80.6, p < .001$), as in Experiment 2. However, unlike in Experiment 2, the full model was also a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(2) = 14.6, p < .001$), indicating a significant polarity \times predictability interaction.

Simple effects were investigated by estimating coefficients for the full model. These are listed in Table 4.6.

Response initiation time

As in Experiment 2, participants generally started to move the mouse within 500 ms of the cursor being released, and often much more quickly, across all conditions, although there was variation in the shape of the distribution by condition (Figure 4.17). Also as in Experiment 2, initiation times were distributed bimodally (Figure 4.17a), as confirmed by a significant Hartigan's dip statistic ($D = .046, p < .001$), with peaks at 39 ms and 306 ms.

Figure 4.18 summarises the variation in response initiation time across conditions. This was modelled across all conditions, with a full set of fixed factors and a random effect of predictability by participant. (Models with a more complex random effects structure failed to converge.) Model comparisons testing for the presence of main effects or interactions were significant for both polarity ($\chi^2(3) = 54.9, p < .001$) and predictability ($\chi^2(4) = 32.9, p < .001$), as in Experiment 2. The full model was also a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(2) = 16.6, p < .001$),

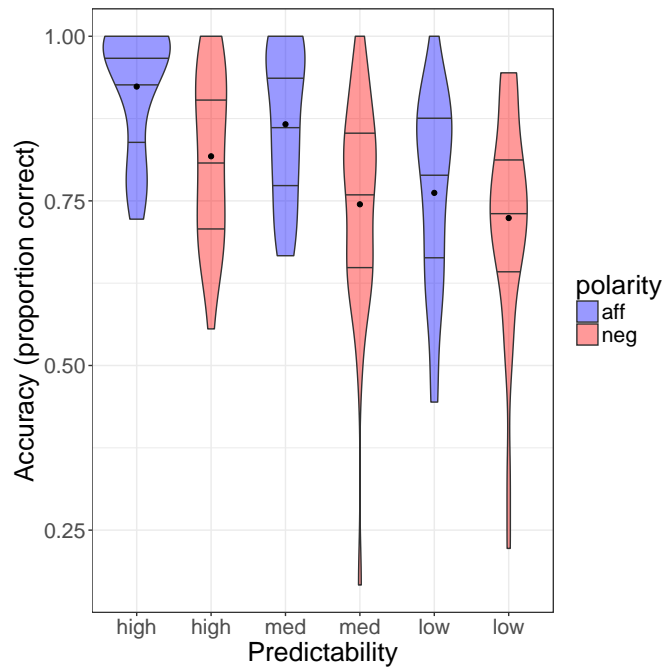


Figure 4.15: Mouse-tracking Experiment 3 response accuracy violin plots: distributions of participant accuracy rates for trials in each condition. Black circles represent means; horizontal lines represent quartiles.

Variable	β	z	p	95% CI	
				lower	upper
Negation					
High pred.	-1.00*	-4.79	< .001	-1.41	-0.59
Med. pred.	-0.82*	-4.65	.001	-1.17	-0.47
Low pred.	-0.19	-1.24	.216	-0.50	0.11
Med. pred.					
Aff.	-0.65*	-3.21	.001	-1.04	-0.25
Neg.	-0.46*	-3.11	.002	-0.76	-0.17
Low pred.					
Aff.	-0.74*	-4.64	< .001	-1.06	-0.43
Neg.	-0.12	-0.84	.402	-0.39	0.16

Table 4.6: Mouse-tracking Experiment 3: simple effects of each factor on the proportion of correct responses at each level of the other factors. For example, the first row provides the effect of negative polarity on the proportion of correct responses in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

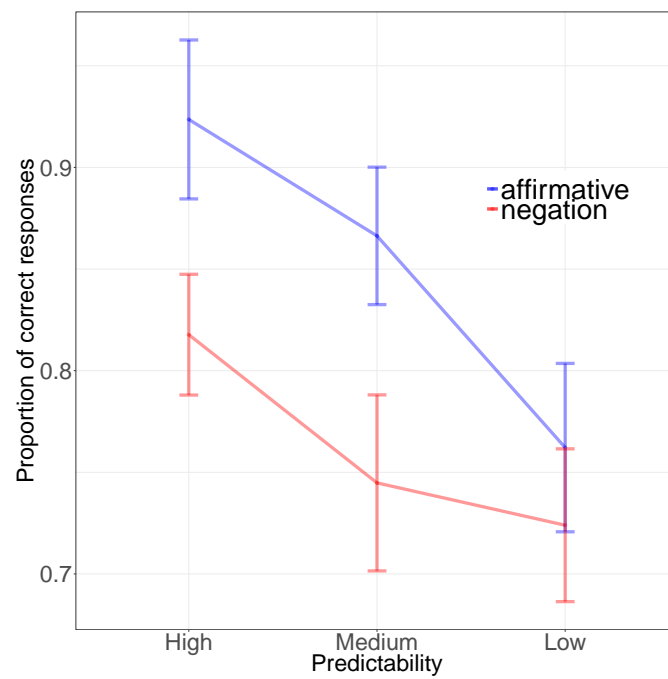


Figure 4.16: Mouse-tracking Experiment 3 response accuracy effects: mean proportion of correct responses across all conditions. Error bars represent 95% confidence intervals.

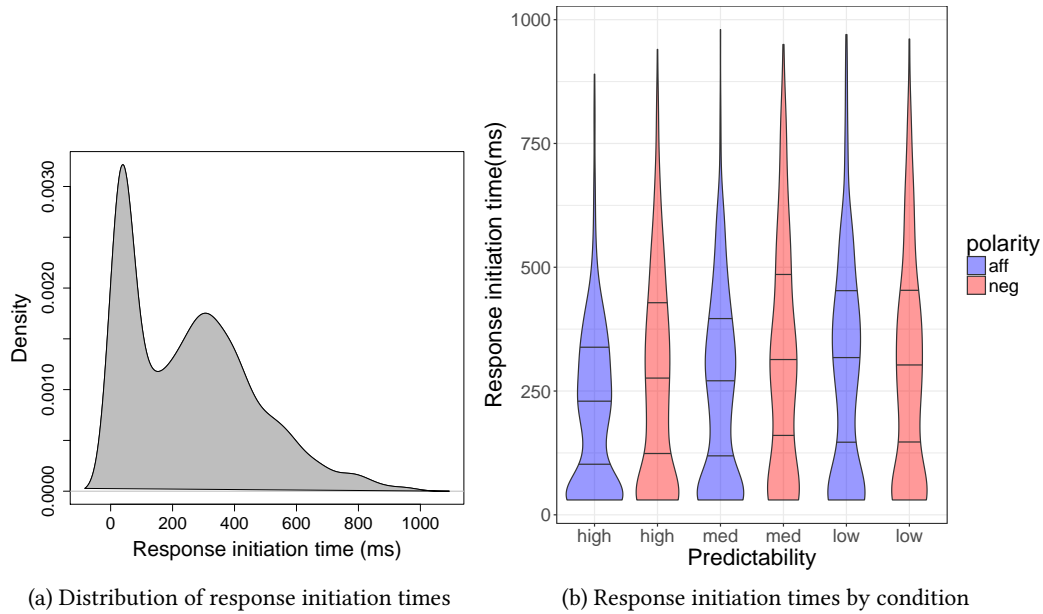


Figure 4.17: Mouse-tracking Experiment 3: distributions of response initiation times, (a) using kernel density estimation, and (b) across conditions. Horizontal lines represent quartiles.

indicating a significant polarity \times predictability interaction, which did not quite emerge in Experiment 2.

Simple effects were investigated by estimating coefficients for the full model. These are listed in Table 4.7.

Response completion time

The distribution of response completion times (after trimming as described above) was slightly skewed (Figure 4.20); although Hartigan's dip statistic suggested significant non-unimodality ($D = 0.026$, $p = .015$), no clear second peak was evident in this case.

Figure 4.19 summarises the variation in response completion time across conditions. This was modelled across all conditions, with a full set of fixed factors and a random effect of predictability included by participant (models with a more complex random effects structure failed to converge). Model comparisons testing for the presence of main effects or interactions were significant for both polarity ($\chi^2(3) = 60.0$, $p < .001$) and predictability ($\chi^2(4) = 62.3$, $p < .001$). The full model was also a significantly better fit to the data than a model including fixed effects

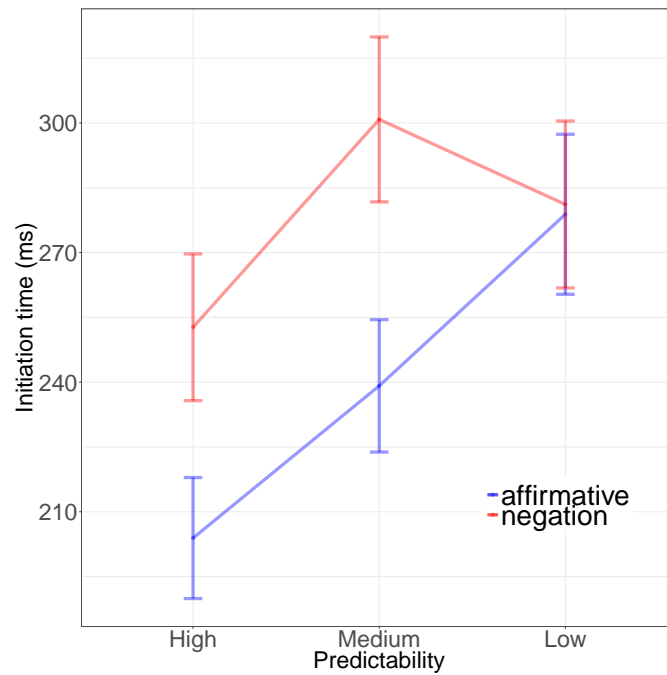


Figure 4.18: Mouse-tracking Experiment 3 initiation time effects: mean response initiation time across all conditions in Experiment 3. Error bars represent 95% confidence intervals.

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Negation						
High pred.	52.52*	4.89	2697.49	< .001	31.46	73.59
Med. pred.	62.99*	5.62	2706.03	< .001	41.03	84.95
Low pred.	1.22	0.10	2713.54	.92	-21.61	24.05
Med. pred.						
Aff.	36.97*	2.77	60.90	.007	10.84	63.11
Neg.	47.44*	3.40	72.33	.001	20.07	74.80
Low pred.						
Aff.	40.43*	3.12	70.33	.003	15.06	65.80
Neg.	-21.34	-1.58	80.59	.117	-47.73	5.05

Table 4.7: Mouse-tracking Experiment 3: simple effects of each factor on response initiation time at each level of the other factors. For example, the first row provides the effect of negative polarity on initiation time in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

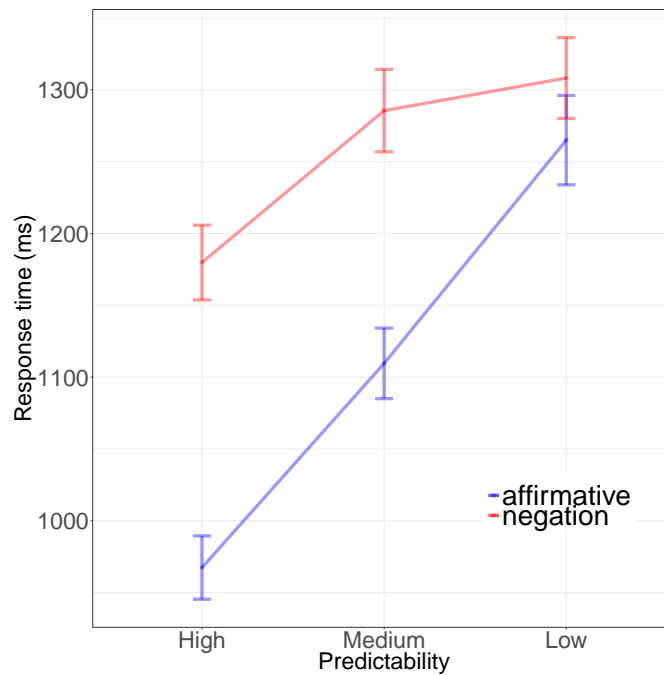


Figure 4.19: Mouse-tracking Experiment 3 response completion time across all conditions. Error bars represent 95% confidence intervals.

of both these factors, but not their interaction ($\chi^2(2) = 22.8, p < .001$), indicating a significant polarity \times predictability interaction. Again, the main effects constituted replications of the preceding experiment, while the interaction did not emerge in Experiment 2.

Simple effects were investigated by estimating coefficients for the full model. These are given in Table 4.8.

Trajectory shapes

As for Experiment 2, the patterns across each dependent measure suggested some level of clustering. Figures 4.21 and 4.22 present all trial trajectories and all data-points sampled throughout the experiment, in further support of this.

This picture is again further supported by clear bimodalities in the AUC and MD, computed for each trial as described above (Figures 4.23 and 4.24). The clustering analysis using $k = 2$ therefore remained a valid approach for this experiment.

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Negation						
High pred.	213.38*	10.04	31.80	< .001	171.74	255.03
Med. pred.	177.04*	7.84	32.17	< .001	132.80	221.28
Low pred.	57.81*	2.14	28.13	.04	4.90	110.71
Med. pred.						
Aff.	143.50*	7.00	35.58	< .001	103.33	183.67
Neg.	107.16*	3.90	29.82	.001	53.32	160.99
Low pred.						
Aff.	156.19*	7.70	43.35	< .001	116.43	195.94
Neg.	36.96	1.58	32.29	.124	-8.88	82.80

Table 4.8: Mouse-tracking Experiment 3: simple effects of each factor on response completion time at each level of the other factors. For example, the first row provides the effect of negative polarity on response time in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

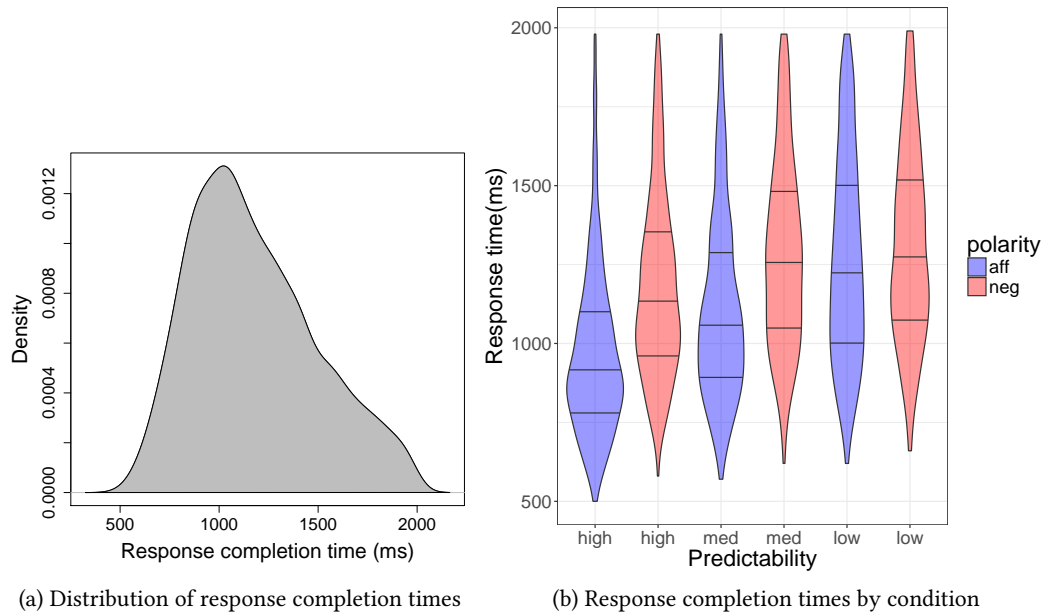


Figure 4.20: Mouse-tracking Experiment 3: distributions of response completion times, (a) using kernel density estimation, and (b) for trials in each condition, measured from release of the cursor. Horizontal lines represent quartiles.

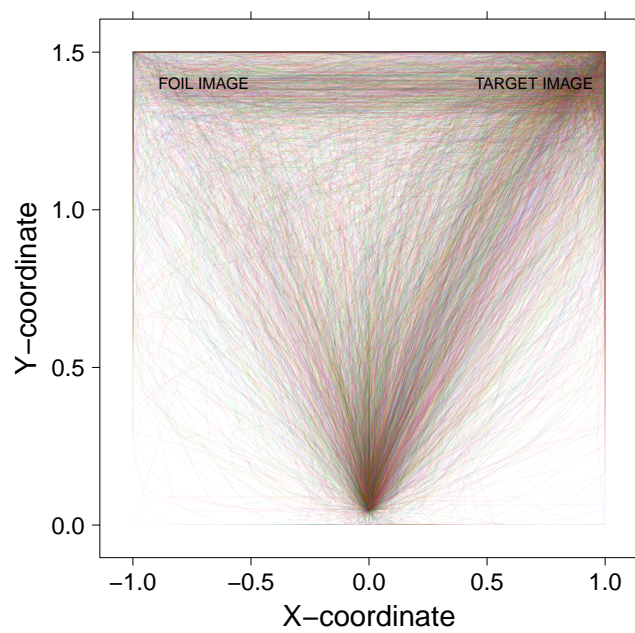


Figure 4.21: All response trajectories in mouse-tracking Experiment 3: a representation of all trials included in the analysis; each line represents the trajectory followed during a single trial, from the starting point in the centre at the bottom of the screen to the finishing point at the target image.

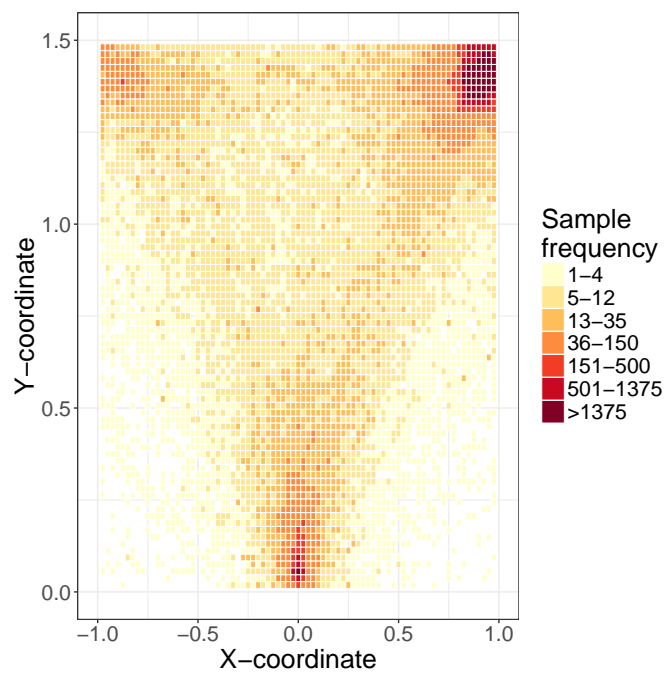


Figure 4.22: Mouse-tracking Experiment 3 heat map of cursor locations: a representation of how frequently, across all trials, any participant's cursor was sampled at each location, using 6,400 (80×80) location bins. Darker colours indicate more common locations. As in Figure 4.21, the starting point is in the centre at the bottom of the screen, the target image in the top right corner, and the foil image in the top left corner.

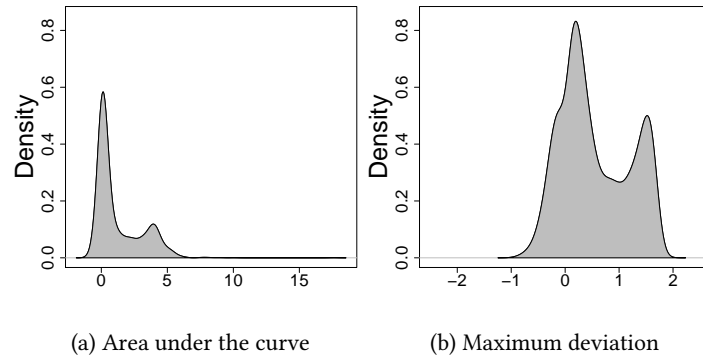


Figure 4.23: Mouse-tracking Experiment 3 area under the curve and maximum deviation: distributions of measures characterising the deviance of individual trials, across all conditions, using kernel density estimation. X axes represent arbitrary units in each case, measuring (a) total area between the trajectory and a straight line between its starting and finishing points, and (b) the distance between the trajectory and the same ideal straight line, at its furthest point.

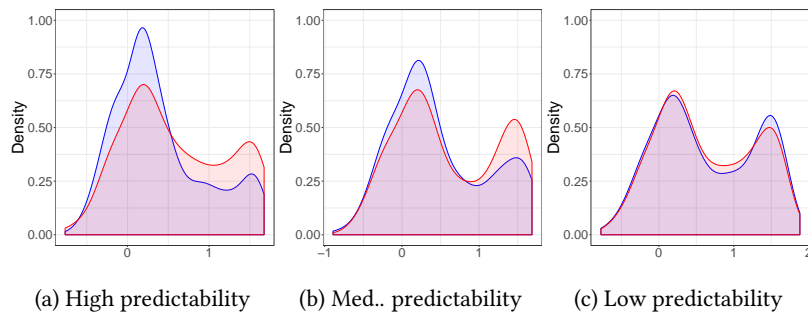


Figure 4.24: Mouse-tracking Experiment 3: distributions of maximum deviation (from a straight line between the starting and finishing points) of trajectories on trials across each condition, using kernel density estimation. Blue plots represent affirmatives; red plots, negations.

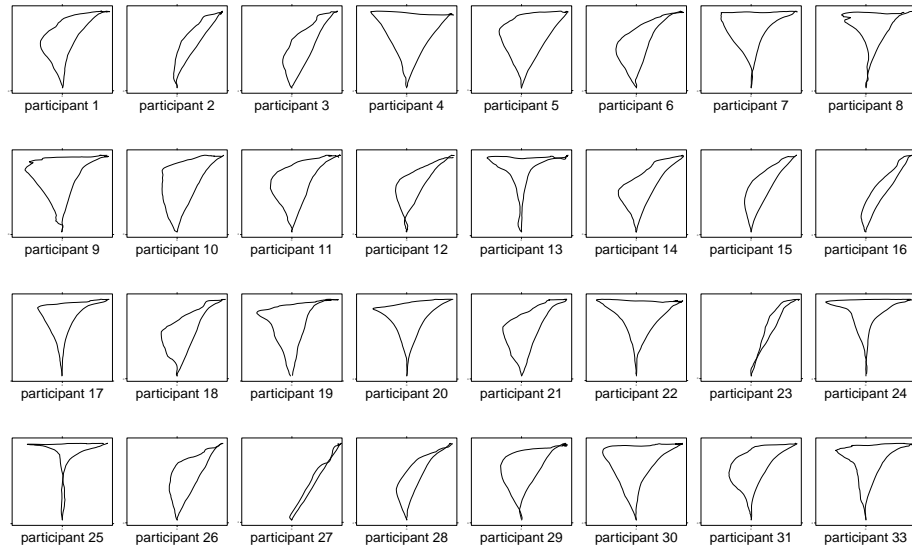


Figure 4.25: Mouse-tracking Experiment 3 individual participant trajectories: the centres of the trajectory clusters identified for each participant in the first stage of the cluster analysis.

Clustering of trajectories

Clustering was carried out as described above. Figure 4.25 illustrates the two cluster centres that emerged for each individual participant; Figure 4.26 shows the results of these cluster allocations combined for all participants.

The proportion of trials falling into each cluster across conditions is shown in Figure 4.27. The proportion of direct-cluster trials was modelled across all participants, with a full set of fixed factors and a random effect of polarity included by participant. (Models with a more complex random effects structure failed to converge.) Model comparisons testing for main effects and interactions were significant for both polarity ($\chi^2(3) = 61.7, p < .001$), and predictability ($\chi^2(4) = 162.0, p < .001$). The full model was also a significantly better fit to the data than a model including fixed effects of both these factors but not their interaction ($\chi^2(2) = 41.3, p < .001$), indicating a significant polarity \times predictability interaction. These findings constituted a replication of Experiment 2 on all counts.

Simple effects were investigated by estimating coefficients for the full model. These are listed in Table 4.9.

To address the concern that more deviant trials simply represent those initiated more quickly, initiation times for trials falling into each cluster were again

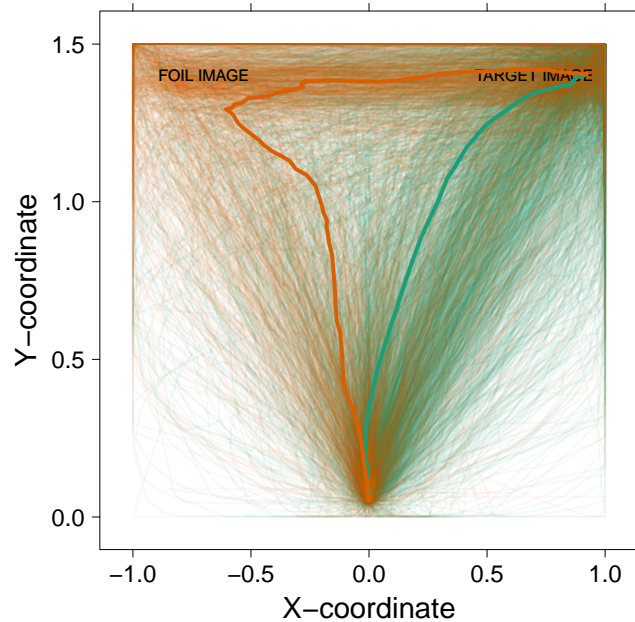


Figure 4.26: Mouse-tracking Experiment 3 trajectory clusters emerging from the cluster analysis. The trajectories of all trials are represented by thin lines (cf Figure 4), colour-coded according to the cluster they fall into. Heavy lines represent the corresponding centres of the clusters.

Variable	β	z	p	95% CI	
				lower	upper
Negation					
High pred.	-1.39*	-6.88	< .001	-1.79	-1.00
Med. pred.	-0.99*	-5.53	< .001	-1.35	-0.64
Low pred.	-0.06	-0.35	.725	-0.40	0.28
Med. pred.					
Aff.	-0.90*	-5.16	.001	-1.25	-0.56
Neg.	-0.50*	-3.53	< .001	-0.78	-0.22
Low pred.					
Aff.	-0.98*	-6.76	< .001	-1.27	-0.70
Neg.	-0.05	-0.35	.728	-0.32	0.23

Table 4.9: Mouse-tracking Experiment 3: simple effects of each factor on the proportion of trials falling into the more direct cluster at each level of the other factors. For example, the first row provides the effect of negative polarity on the proportion of direct-cluster responses in the high predictability condition. The reference levels of each factor are high (for medium predictability), medium (for low predictability), and affirmative (for negative polarity).

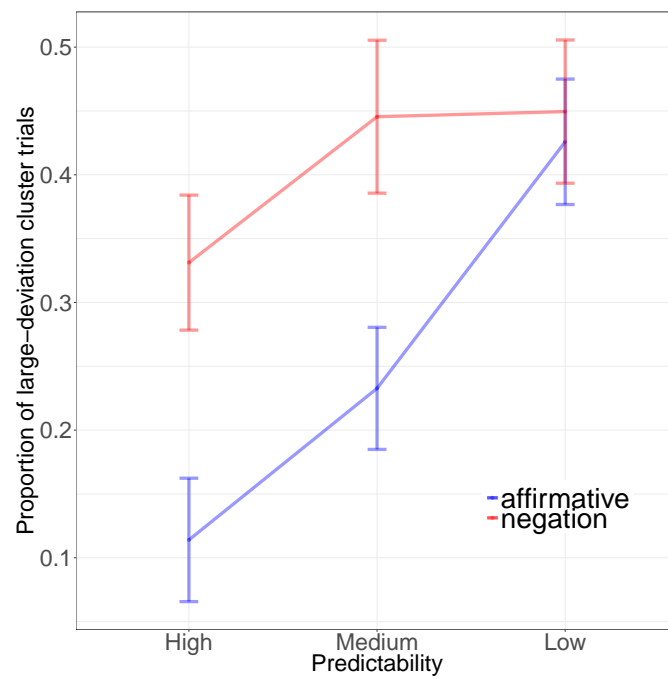


Figure 4.27: Mouse-tracking Experiment 3 effects on trajectory cluster: the mean proportion of trials falling into the more deviant cluster across each condition. Error bars represent 95% confidence intervals.

compared. As for Experiment 2, the mean initiation time was greater for trials falling into the more foil-skewed cluster, and a paired samples t test showed that this difference was significant ($t(31) = -6.09$, $p < .001$). To reiterate, this check is important because, if initiation times were longer for trials falling into the more direct cluster than for those falling into the more foil-skewed cluster, the evidence would suggest that cluster shape per se could not tell us much about processing during the trial; instead, it would seem to be the case that participants simply went directly to the correct answer on trials where they waited long enough to initiate their response to have figured this out, or took a quick guess (resulting half the time in a mid-trial change of mind) early on. Because the opposite pattern is observed, this bolsters confidence that trials falling into the more foil-skewed cluster represent something more than guessing: arguably, at least in some cases, a positive attraction to the foil.

4.4 Discussion

Experiments 2 and 3 both replicated the consistent main effects observed across all dependent variables in Experiment 1. That is, both negative polarity and decreasing predictability consistently decreased the accuracy of participants' responses, making them slower to initiate and complete their responses, and more likely to exhibit attraction to the foil response. This constitutes further evidence supporting the conclusions arising from the main effects found in Experiment 1: namely, that negation can have a detrimental effect on processing even for the type of pragmatically licensed sentences presented in these experiments, and that reducing the predictability of the critical word or sentence completion (as expected) makes the task more difficult.

However, the findings in terms of the interaction between polarity and predictability were very different to the weak and unclear evidence arising from Experiment 1. Experiment 2 also presented relatively weak evidence for an interaction, which was statistically significant only in the cluster analysis (although this is the metric capturing the most relevant information about participants' responses). However, this interaction was in the opposite direction to that observed in Experiment 1. While the Experiment 1 findings tended to suggest (in line with the original hypothesis) that reducing predictability exacerbated the processing difficulties imposed by negation, this conclusion was not supported by Experiment 2 and indeed was contradicted by Experiment 3. In this case, where responses were

	Experiment 1	Experiment 2	Experiment 3
Response accuracy	<ul style="list-style-type: none"> • detrimental main effect of negative POL, stronger for true than false sentences • reducing PRED affected true and false sentences inconsistently depending on polarity • 3-way $POL \times PRED \times TV$ interaction 	<ul style="list-style-type: none"> • detrimental main effect of negative POL except at low PRED • detrimental main effect of reducing PRED only from high to medium • no interaction 	<ul style="list-style-type: none"> • detrimental main effect of negative POL except at low PRED • detrimental main effect of reducing PRED consistent except from medium to low for negatives • $POL \times PRED$ interaction
Response initiation time	<ul style="list-style-type: none"> • detrimental main effect of negative POL except for medium PRED false sentences • detrimental main effect of reducing PRED (inconsistent across conditions) • no main effect of TV • no 3- or 2-way interactions • no simple interactions 	<ul style="list-style-type: none"> • consistent detrimental main effect of negative POL • detrimental main effect of reducing PRED only for affirmatives • no interaction (perhaps due to underpowering) 	<ul style="list-style-type: none"> • detrimental effect of negative POL except at low PRED • detrimental effect of decreasing PRED except from medium to low for negatives • $POL \times PRED$ interaction
Response completion time	<ul style="list-style-type: none"> • detrimental main effect of negative POL • detrimental main effect of reducing PRED (disappears for low PRED affirmatives) • detrimental effects of falsity for all affirmative but no negative sentences • 2-way $POL \times TV$ interaction with beneficial effect at all levels of PRED 	<ul style="list-style-type: none"> • consistent detrimental main effect of negative POL • almost consistent detrimental main effect of decreasing PRED • no interaction 	<ul style="list-style-type: none"> • detrimental effect of negative POL reduced at low PRED • detrimental effect of decreasing PRED except from medium to low for negatives • $POL \times PRED$ interaction

Cluster analysis of trajectory shapes	<ul style="list-style-type: none"> • detrimental main effect of negative POL • detrimental main effect of reducing PRED (only consistent from high to medium) • detrimental main effect of false TV except for high PRED negations • no interactions 	<ul style="list-style-type: none"> • detrimental main effect of negative POL weaker at lower PRED • detrimental main effect of reducing PRED weaker for negations • POL \times PRED interaction 	<ul style="list-style-type: none"> • detrimental main effect of negative POL except at low PRED • detrimental main effect of reducing PRED reduced or disappearing for negatives • POL \times PRED interaction
Summary	<ul style="list-style-type: none"> • mixed findings present an unclear picture • effects of TV mask some of the potential effects of interest • where patterns do emerge they suggest a benefit of higher PRED in processing negative sentences 	<ul style="list-style-type: none"> • clear detrimental main effects of both negative POL and lower PRED • no interaction conclusively emerges 	<ul style="list-style-type: none"> • a much more consistent and interpretable set of results • main effects of POL and PRED are accompanied by a clear interactions • simple effects show that this consistently goes in the opposite direction to the hypothesis: the detrimental effects of negative POL are smaller at lower PRED

Table 4.10: A summary of the main findings across all three mouse-tracking experiments. The abbreviations POL, PRED and TV refer to the independent variables of polarity, predictability and truth value, respectively.

speeded (and more participants tested), the picture was much clearer: a polarity \times predictability interaction strongly emerged, present across all dependent variables and in the same direction as the weak pattern found in Experiment 2, with reducing predictability eliciting more similar results for affirmatives and negatives, rather than a bigger difference.

The overall findings for each dependent variable across Experiments 1, 2, and 3 are summarised in Table 4.10.

4.4.1 Was there a floor effect?

In interpreting this interaction, one concern is that it may arise from a floor effect. That is, perhaps performance for negations was already so poor even on the easiest, high-predictability trials that there was little room for it to fall further, whereas performance for affirmatives was good enough in the easier conditions to allow for a large drop in performance when predictability decreased. The lack of change between medium- and low-predictability conditions for negations in Experiment 3 might be regarded as particularly indicative of this.

However, this interpretation is only reasonable if it is the case that no more than approximately 50% of trials should ever fall into the large-deviation cluster. This would be true if, in the case of the worst possible performance, allocation of a trial to a cluster were based purely on guessing: an initial correct guess would correspond to a trial falling into the more direct cluster, whereas an initial incorrect guess would correspond to a trial falling into the foil-skewed cluster, and a guess is equally likely to be correct or incorrect, meaning that approximately 50% of trials would fall into each cluster. Although the conditions with the worst performance in both Experiments 2 and 3 do show approximately 50% of trials falling into each cluster, there are two main indications that this is probably not the result of a floor effect. The second is discussed below in relation to how foil-skewed cluster trials should be interpreted (whether as guesses, or as active attraction to the foil response); here, I note simply that the proportion of trials falling into foil-skewed cluster in this type of paradigm can exceed 50% in reality as well as in principle, as it did in Experiment 1 (see Figure 3.15).

4.4.2 Anti-prediction

Another concern in interpreting the data, as mentioned in the Introduction to Chapter 3 (section 3.1.3), is whether participants' use of an "anti-prediction" strat-

egy could be a possible explanation for the pattern observed. The availability of this strategy only arises in a very specific case (high-predictability negations), and even then only if participants do *not* process negation incrementally and therefore cannot update their predictions on this basis. That is, if participants' initial predictions are consistently wrong for negations, *and* they make predictions of the form "not A" in preference to "B, C, or D", then these effects could cancel out, making participants' performance on high-predictability negations appear rather strong even though this results from compounding of two mistakes. This interpretation is compatible with the trajectory data observed in Experiment 3 (and perhaps, to a much weaker extent, in Experiment 2), where the decrement in performance between high- and medium-predictability negations is much larger than the decrement between medium- and low-predictability. However, two pieces of evidence indicate that anti-prediction is an unlikely explanation for this pattern.

First, in the case of both Experiments 2 and 3, the slope for affirmatives in the trajectory cluster analysis between high and medium predictability is roughly the same as the slope for negations. Anti-prediction was not an available strategy for critical affirmative trials, so the slope for affirmatives can be attributed solely to the change in predictability. This shows that predictability can impose at least this magnitude of effect, without the involvement of anti-prediction.

Second, and rather more convincingly, the appeal of the anti-prediction strategy to participants can be investigated by looking at affirmative trials where it was available. This was not the case for any critical trials, but for some counterbalancing (constant-predictability) affirmative trials (those associated with 4-item images), there was a single possibility for the foil response option, and thus the anti-prediction strategy was available. If participants preferred to use it in the case of unincrementally-processed negations, they should also prefer to use it in this case, where it should be in evidence in the form of a detrimental effect on their performance (because the priming in this case would direct them initially towards the wrong answer). Thus, participants' performance on these counterbalancing trials can be compared to their performance on low-predictability critical affirmative trials. (All counterbalancing trials were equivalent to low-predictability critical trials, as there were three possible items for the target response option.) In Experiment 2, 30% of the relevant affirmative counterbalancing trials fell into the more deviant cluster of trials, as opposed to 38% of the comparable critical trials, although a paired samples *t* test showed that the difference did not reach significance, $t(23) = -2.07$, $p = 0.049$. In Experiment 3, the figures were 28% and

43%, respectively, with the difference this time reaching significance ($t(31) = -5.17$, $p < .001$).

A finding in this direction cannot absolutely rule out anti-prediction, because the comparison is imperfect: the critical affirmative trials contained more objects that could play the role of the foil, even though predictability was the same, so a poorer performance in this condition might have arisen from the increased memory load even if performance was affected in the counterbalancing condition by interference from the anti-prediction strategy. However, for this to be the case, the effect of the anti-prediction strategy would have to be extremely small, in order to be completely overwhelmed by a memory load effect producing an overall large difference in performance (at least in the case of Experiment 3) in the opposite direction. This finding can therefore be taken as weak evidence that participants did not use the anti-prediction strategy, or that if they did so, its impact on the data was probably very small, particularly in Experiment 3.

4.4.3 Interpreting the findings

Having considered these potential nuisance interpretations of the data (concluding that the pattern can be best explained primarily neither by floor effects nor by the use of an anti-prediction strategy by participants), I turn to the main question of what it can tell us about predictability, polarity, and their interactive influence on sentence processing.

The relative roles of predictability and memory load

One important question to consider is to what extent the manipulation of predictability in Experiments 2 and 3 isolated this factor specifically. Because predictability was manipulated by adjusting the total number of objects in the visual display, predictability was potentially confounded with the total memory load on participants, and some of the main effect of predictability can probably be attributed to the increase in task difficulty on this basis. Storing multiple objects for later recall in prediction making represents a drain on cognitive resources, and as well as contributing to the main effect of predictability, this memory load effect could in itself be responsible for some of the apparent predictability \times polarity interaction.

The constant-predictability trials presented for counterbalancing purposes were used to test this. All of these trials were effectively “low-predictability” trials,

as the correct answer could be one of three options regardless of whether there were four, five or six objects in the grid in total. Thus, rather than differing in predictability of the available correct response option, they differed in how many potential foils there were: one, two, or three. Participants' performance could therefore be compared between counterbalancing and critical trials that contained the same number of objects. As the number of objects increases, any degradation in performance in the counterbalancing trials can be attributed solely to the increase in memory load; thus, the interaction between the number of objects and whether the trial is a constant-predictability one (i.e., an original counterbalancing trial) or a varying-predictability one (i.e., an original critical trial) should provide a measure of the effect of predictability, as disentangled from memory load. This analysis was carried out for the proportion of trials falling into each cluster in Experiments 2 and 3.

As illustrated in Figure 4.28, the slopes differed between varying-predictability and constant-predictability lines, showing that there was an effect of predictability over and above the increased memory load, but while this effect was clearly present in affirmatives (in both Experiments 2 and 3), it was weaker or absent in negations. To examine this interaction, mixed effects logistic regression models (as used in the main analysis of results) were constructed separately for affirmatives and negatives in each experiment to model the proportion of trials falling into each cluster, including fixed factors of predictability (constant or varying), number of objects present in the display (4, 5, or 6), and the interaction between these two factors. A random effect of predictability by participant was also included. (Models with a more complex random effects structure failed to converge.)

For affirmative trials in Experiment 2, model comparisons testing for the presence of main effects or involvement in an interaction were significant for both predictability ($\chi^2(2) = 23.02, p < .001$) and number of objects ($\chi^2(2) = 75.70, p < .001$). The full model also represented a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(1) = 15.66, p < .001$), indicating a significant predictability \times number of objects interaction. For negative trials in Experiment 2, model comparisons testing for the presence of main effects or involvement in an interaction were also significant for both predictability ($\chi^2(2) = 9.37, p = .009$) and number of objects ($\chi^2(2) = 58.08, p < .001$). However, in this case the full model did not represent a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(1) = 2.35, p = .13$), indicating a lack of a significant

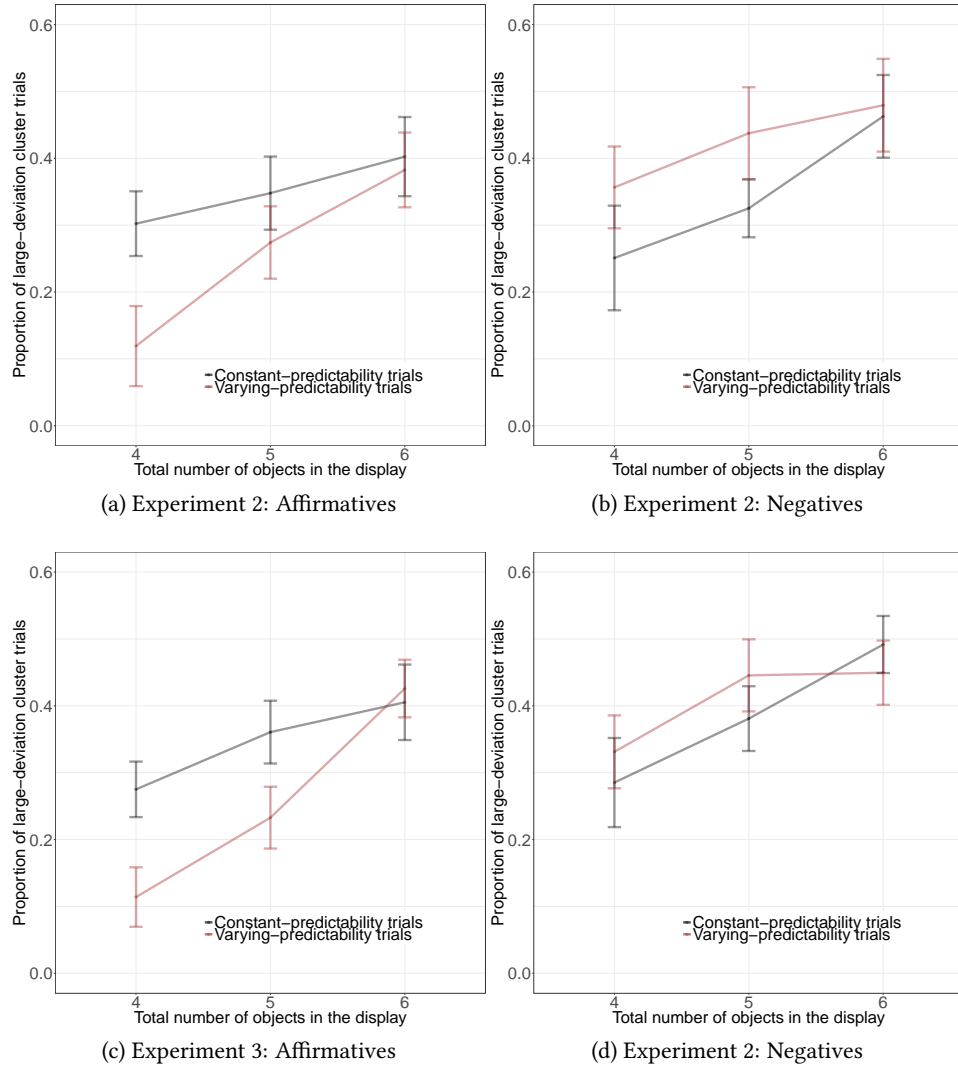


Figure 4.28: Cluster allocations for varying-predictability and constant-predictability trials in Experiments 2 and 3: comparisons of the proportion of trajectories falling into each cluster. The lines representing varying-predictability trials (original critical trials) are identical to those seen in Figures 4.14 and 4.27. The slopes of the constant-predictability trials represent the effects of introducing additional objects, while keeping predictability constant; the difference between these slopes and those of the varying-predictability trials represent the effects of predictability, over and beyond the increased memory load resulting from additional objects.

predictability \times number of objects interaction.

For affirmative trials in Experiment 3, model comparisons testing for the presence of main effects or involvement in an interaction were significant for both predictability ($\chi^2(2) = 48.33, p < .001$) and number of objects ($\chi^2(2) = 163.29, p < .001$). The full model also represented a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(1) = 33.86, p < .001$), indicating a significant predictability \times number of objects interaction. For negative trials in Experiment 3, model comparisons testing for the presence of main effects or involvement in an interaction were significant for number of objects ($\chi^2(2) = 64.01, p < .001$), but not for predictability ($\chi^2(2) = 4.85, p = .088$). The full model also represented a significantly better fit to the data than a model including fixed effects of both these factors, but not their interaction ($\chi^2(1) = 4.30, p = .038$), indicating a significant predictability \times number of objects interaction.

The overall pattern emerging from this analysis is that the effect of predictability over and beyond the associated effect on memory load in critical trials was much stronger in both experiments for affirmatives than for negations. In the case of negations, the interaction between predictability and number of objects was not present at all in Experiment 2 and was rather weak in Experiment 3. Thus, although some of the effects of predictability can be attributed to the association with memory load (especially in the case of negative sentences), memory load cannot be responsible for the polarity \times predictability interaction observed in the main analysis; in fact, the effect of memory load was fairly stable across affirmatives and negations and fairly linear in the total number of objects presented, thus only exaggerating somewhat the size of the main effect of predictability and interfering very little with the polarity \times predictability interaction. It can be concluded that, as suggested by the original comparisons (especially in Experiment 3), reducing predictability does have a more detrimental effect overall on affirmatives than on negatives, the underlying cause of which is not an increase in memory load arising from the experimental design.

Interpreting the interaction between polarity and predictability

Particularly in Experiment 3, the polarity \times predictability interaction arises primarily from the disparity between affirmatives and negations in the shift from medium to low predictability, which induces a large change in the proportion of trials falling into each cluster in affirmatives, and a very small or entirely ab-

sent change in negatives (see Figure 4.27). As discussed in section 4.4.1 above, a floor effect on performance for negations is not the most likely explanation for this pattern. Rather than attempting to explain the relatively weak effect of predictability on negatives, it may be more enlightening to explain the relatively *strong* effect on affirmatives of increasing predictability relative to a low default. That is, this pattern could be interpreted as representing a superadditive effect of predictability and polarity, in which affirmative polarity and high (or medium) predictability each do little in isolation to lift performance away from a baseline at low predictability, but their combination (in the affirmative, high-predictability condition, and to a lesser extent the affirmative, medium-predictability condition) has a powerful beneficial effect on performance.

Indeed, it is intuitively reasonable to imagine that prediction-making is an easy and appealing strategy for affirmative sentence fragments with only one possible completion, or perhaps with two, but the addition of any further complexity (in the form of negation or reduced predictability) relatively easily obliterates this ease of the task, rendering all other conditions effectively equivalent. Conceivably, participants make trial-by-trial decisions (whether consciously or unconsciously), based on the global complexity or resource consumption of the overall trial conditions, as to whether to attempt to make a prediction at all.

Interpreting foil-skewed trials

The above interpretation of the data suggests merely that prediction-making breaks down in sufficiently complex conditions; this fails to provide any information specifically about the incrementality (or lack thereof) of negation processing. The same conclusion could be drawn whether poor performance results from guessing or from active attraction to the foil response option (which would indicate initially non-incremental processing, arriving at a positive initial prediction for the foil).

At first glance, guessing might seem the most likely explanation for conditions with a high proportion of deviant cluster trials, especially because the affirmative condition (in which there is no reason for attraction to the foil response to occur based on a lack of incremental processing) give rise to just as many large-deviation trials at low predictability as the negative condition. However, closer examination of the pattern of results obtained arguably leads to the conclusion that at least some trials in some conditions fall into the more deviant cluster as a result of attraction to the foil, rather than simple guessing.

First, as noted in the Results (sections 4.3.3 and 4.3.4), trials falling into the

more deviant cluster were, on average, initiated more slowly than those falling into the less deviant cluster. The opposite pattern would be expected if highly deviant trials arose from time-pressured participants initiating their mouse movements quickly with a random guess.

Furthermore, because quickly-initiated trials were more likely to fall into the more direct cluster, this cluster can be taken to represent trials where the participant could quickly discern and execute the correct response. If an assumption is made that all or most of the rest of the trials (those initiated more slowly) represent cases in which the participant was forced to take an initial guess, after delaying their response for too long and still not yet having arrived at an answer, these trials would be equally likely to head initially towards the foil or the target, and thus would be approximately equally likely to fall into each cluster. Therefore, if this assumption were true and such trials were the only ones in the experiment, about 50% of trials would fall into each cluster. Of course, such slow-initiation trials were not the only ones in the experiment; however, the remaining fast-initiation trials (as already established) tended to fall into the more direct cluster, driving the total proportion of foil-skewed trials (under the aforementioned assumption) down rather than up. Therefore, given that the proportion of foil-skewed trials does in fact approach 50% in some conditions, the only way the assumption that slow trials represent guessing can hold would be if the proportion of fast-initiation trials (increasing the size of the direct cluster) were very small. This can be checked by counting how many trials altogether fall under the early peak of the bimodal initiation time distribution: the proportion of trials with a faster initiation time than the lowest point of the dip in the distribution between the two peaks was 28% in Experiment 2, and 39% in Experiment 3. These are substantial proportions, indicating that something beyond guessing must be in operation in order for the proportion of large deviation cluster trials to approach 50%.

Overall, the results suggest that non-incremental processing is likely to be responsible for at least some of the attraction towards the foil response option for negations. However, it remains unclear how to explain the very similar levels of attraction towards the foil in the case of the low-predictability affirmative conditions.

4.4.4 Conclusions

Both Experiments 2 and 3 showed a similar pattern of results. The main effects of polarity and predictability identified in Experiment 1 were replicated; however, in

contradiction of the hypothesis, the polarity \times predictability interaction was in the opposite direction to the weakly evidenced interaction seen in Experiment 1. That is, decreasing predictability had a weaker effect on negatives than on affirmatives.

Although the broad pattern of results was the same in Experiments 2 and 3, the findings were not identical. In particular, Experiment 2 exhibited the same trends as Experiment 3 in terms of the interaction, but this reached significance only for the cluster analysis of trajectories. The experiments were extremely similar, differing only in the amount of pressure exerted on participants to respond quickly (which was greater in Experiment 3) and statistical power (as data was collected from more participants in Experiment 3). It thus seems reasonable to conclude that the underlying effects were the same across both experiments; it was simply easier to detect the polarity \times predictability interaction in Experiment 3 because of the increased power and the increased access to early cognition provided by the additional pressure on participants to initiate their response quickly.

Not only does this set of results suggest that the “active ingredient” in pragmatic felicity, as it interacts with negation, is not predictability (as hypothesised), it is also inconsistent with Nieuwland and Kuperberg’s (2008) finding that pragmatically licensed affirmatives and negations are processed equivalently easily; instead, a strong main effect of negation was observed. It is possible that this can be attributed to methodological differences, although Dale and Duran (2011) obtained results consistent with those of Nieuwland and Kuperberg (2008) using a mouse-tracking paradigm. A more likely explanation seems to lie in the exact nature of the sentences (or sentence fragments) and the contexts in which they were presented in the present experiments.

In particular, although the sentences used in these experiments were licensed by the context, this licensing arose through the nature of the general task (a kind of memory game), rather than through rich contextual information that might, for example, activate particular semantic fields or license the use of a negation *specifically* to deny or contradict an assertion or assumption present in the discourse. In this way, the sentences presented in the experiments described here may have been most similar to the “simple context” condition in Dale and Duran (2011), as opposed to their pragmatically richest condition, the latter being the case in which their results followed the same pattern as Nieuwland and Kuperberg (2008). This is not to argue that the stimuli used here were improperly licensed, but a distinction between licensed and fully contextually felicitous may be of importance in this context. It is also important to note that because the same broad

framing was the source of the licensing for all sentences in these experiments, caution is required in generalising the findings to conclude that the effects would hold in different types of licensing context, including more naturalistic ones.

Finally, although these experiments have not provided a firm understanding of the mechanism underlying the relationship between pragmatic felicity and negation, they have highlighted some clear advantages and disadvantages of certain methodological approaches over others. In particular, the truth-value judgement paradigm interacts inconveniently with manipulations of polarity, as well as adding a superfluous layer of complexity and task difficulty when the latent process of interest is prediction-making. The sentence completion task provides much clearer and less noisy access to the content of participants' predictions in this type of experiment.

Chapter 5

Eyetracking Study

This study was designed and carried out during a visit to the Linguistics department at the University of Maryland in autumn/winter 2015. Colin Phillips contributed greatly to the concept and design of the experiment and discussions with a number of other UMD lab members were also helpful. Julia Buffinton assisted with practical lab matters (along with Anton Malko), supplied audio recordings of the materials, and conducted several of the data collection sessions. Jon Burnsky and Hanna Muller subsequently took over the project and have conducted several follow-up experiments (briefly described in the Discussion).

5.1 Introduction

5.1.1 Background

As identified in the course of the preceding experiments, clearer results in paradigms aiming to measure incremental processing of linguistic input can be obtained by tapping more directly into participants' cognitive processes. Two strategies to achieve this are to access richer information on earlier stages of processing, and to use behavioural measures more directly related to the predictions made by comprehenders. In this context, eye-tracking is evidently a methodology that merits consideration. As with hand-movements made in the process of supplying a response (exploited in mouse-tracking), people's eye movements have been found to reflect their cognition: specifically in relation to language processing, individuals tend to direct their gaze towards a visual depiction of an object as soon as it is referred to in an accompanying sentence (e.g. Allopenna, Magnuson, & Tanenhaus, 1998).

This relationship between eye movements and language processing has been put to extensive use in studying the incrementality of language processing. For example, Altmann and Kamide (1999) presented quasi-realistic visual scenes and auditory sentences in which the verb either picked out a certain entity in the scene as most likely to function as the object of the verb, or could take one of multiple available objects (“The boy will *eat/move* the cake”). Participants shifted their gaze sooner to the location of the object referred to in the case of the restrictive verb than in the case of the non-restrictive verb, illustrating their incremental incorporation of information on the semantics of the verb and its impact on upcoming material in the sentence. Kamide et al. (2003) showed using a similar approach that this principle extends to other types of syntactic and semantic processing and prediction: for instance, the verb can also be used to predict an indirect object appearing later in the sentence (e.g., a goal rather than a direct object); an agent preceding the verb can contribute to the prediction of a theme following the verb; and in a head-final construction (in Japanese), the verb can be constrained by arguments appearing before it. Other work has explored the extent to which various types of linguistic and paralinguistic information can be used to update predictions incrementally (e.g. Sedivy et al., 1999, on scalar implicatures of scalar adjectives).

Various experimental paradigms exploring negation have made use of the fact that visual stimuli can be paired with linguistic input to investigate which representations are primed or activated to a greater extent at different stages of processing (e.g. Kaup et al., 2006; Lüdtke et al., 2008; Tian et al., 2010). However, these approaches tend to make use of accuracy or response time based measures; relatively few studies have used visual world-type paradigms accompanied by eye-tracking techniques to investigate the processing of negative sentences.

Orenes et al. (2014) employed a visual world paradigm to examine how negation is conceptualised in a “binary” as compared to a “multary” context. They presented a visual world consisting of four coloured shapes, preceded by a binary (“The figure could be red or green”) or multary (“The figure could be red, green, blue or yellow”) context sentence. The image was then accompanied by an affirmative or negative sentence of the form “The figure is / is not red”. In the affirmative case, participants rapidly (within about 400 ms of the onset of the adjective in the critical sentence) shifted their gaze towards the shape of the mentioned colour, regardless of whether the context provided was binary or multary. In the negative case, participants looked at the alternative (in this case, the green shape) in the

binary condition, where there was a defined alternative to the mentioned colour. However, in the multary context, they were more likely to look at the shape of the mentioned colour than at any of the alternatives, even though it was the only one explicitly ruled out in the description. The authors interpret the overall pattern as evidence that negation can be interpreted symbolically — that is, as a mental “tag” attached to a specific concept or entity — in cases where the negation does not combine with the negated entity to form a representation of an alternative.

Orenes et al. (2016) followed this experiment with an eye-tracking study on the effects of the felicity of affirmative and negative sentences, as manipulated by contextual information. They presented displays containing images representing opposing concepts or entities (e.g., a rich man and a poor man) and measured looks to each of these images in several conditions. Linguistic stimuli consisted of two sentences. The second picked out one of the images as a target, either in an affirmative or a negative form (e.g., either “Her dad was rich” or “Her dad was not poor” rendered the image of the rich man the target). Preceding this sentence, a context sentence provided information that was consistent (“She supposed that her dad had enough savings”), inconsistent (“She supposed that her dad had little savings”), or neutral (“Her dad lived on the other side of town”) with respect to the critical sentence. They found that looks to the target image increased earlier for negative sentences preceded by an inconsistent context sentence compared to a neutral context sentence, whereas participants took longer to look at the target in the case of negative sentences preceded by a consistent context sentence. This was taken as evidence that the inconsistent pragmatic context pre-activated the concept (in this case, *poor*, as embodied in the image of the poor man) that would subsequently be negated in the critical sentence, meaning that only one further step was required for interpretation (from *not poor* to *rich*); in contrast, in the consistent case, the opposite context was activated, meaning that an additional step was required (from *rich* to *not poor* back to *rich*). However, looks to the target image increased sooner when the critical sentence was affirmative than when it was negative, regardless of the context sentence, suggesting that negation always incurs a processing difficulty, even when the pragmatic context is facilitatory. This pattern could explain why pragmatic context is often found to be beneficial for the processing of negation, as negative sentences are more readily interpreted when they are presented as a contrast to or denial of a previously asserted or assumed state of affairs.

The latter experiment specifically related to binary or “bipolar” predicates, i.e.

those with a specific opposite (*rich* vs. *poor*); as pointed out in previous chapters, processing of negation is often facilitated in this type of case, as there is a ready-made contrast frame (Löbner, 2000) directing interpretation of the negation to a specific meaning. Combined with the preceding pragmatic context, this also makes such sentences highly predictable, reducing the need for a “search for alternatives” (Ferguson et al., 2008). However, as in previous studies manipulating pragmatic felicity, it is difficult to use similar paradigms to explore the impact of episodic predictability while controlling for pragmatic felicity.

5.1.2 The present study

The present study adopted a preliminary approach, aiming to investigate the extent to which participants’ gaze in a visual world paradigm could be manipulated by the use of affirmative and negative sentences involving descriptions of the visual world. The experiment was designed with a view to building more complex paradigms on this basis, to allow for further investigation of the factors (including predictability) that would affect participants’ ability to incrementally update their predictions about upcoming material in the course of processing an incoming negative sentence.

To achieve this preliminary aim, a design was constructed in which participants heard sentences describing the location of a target object in a grid presented visually (a similar approach to the designs used in Chapters 3 and 4). They could incrementally use the information in the sentences (which took the form “The left side contains the cactus. It also contains / doesn’t contain the pitcher”) to narrow down the possible candidates for a target object, which would eventually be identified uniquely, at intermediate points in the sentence. In the case of a negative sentence (“doesn’t contain”), participants would need to interpret the negation in order to make the correct inference about the possible candidates; waiting for the full proposition in order to interpret the negating element would lead to predicting that foil objects in the visual world (in this case, those in the leftmost column) would be good candidates for the target, leading to looks to the foils rather than possible candidates.

As well as the critical sentence, this design also included context sentences (as in the example above). The context sentence preceded the critical sentence containing the name of the target object. The purpose of these was twofold. First, it was desirable to equate the timing and structure of the affirmative and negative critical sentences as closely as possible. The use of a context sentence enabled the

felicitous inclusion of the word *also* in the affirmative sentences in place of *doesn't*, which was required in the negative sentences. Second, the initial sentence provided a context in which both the affirmative and negative sentences were pragmatically licensed. Previous work (Dale & Duran, 2011; Nieuwland & Kuperberg, 2008, etc.) has shown that pragmatic felicity is an important condition to enable negative sentences to be processed incrementally as readily as alternatives; in this case, the negative sentences formed a specific, relevant contrast with an affirmative presented in the immediately preceding sentence, ensuring that participants would not interpret these negations as unlicensed or appearing “out of the blue” when there is no particular reason to deny or contradict a proposition that was not apparently under discussion.

5.2 Methods

5.2.1 Materials

Stimuli

Materials consisted of images and accompanying auditorily presented sentences, with the images used as episodically-presented contexts for the sentences in a similar way to the mouse-tracking experiments (Chapters 3 and 4). Each image consisted of a 3×3 grid, five cells in which were occupied by a picture of an everyday object. These were drawn from a set of 120 such objects, selected for consistent naming in American English.

The objects were arranged in the grid in a configuration such that one row or column was completely filled and another had a single empty cell (see Figure 5.1 for an example). The middle row or column was always left empty. Thus, one item was always an “odd one out”, alone in its column or row. Another of the items (not the “odd one”) functioned as the target.

Each set of five objects was used to produce two images, identical except for an exchange in position and role of the items in the target column or row (for counterbalancing purposes). Each image was also associated with two possible auditory inputs (affirmative and negative). In both cases, the auditory input began with the same context sentence, mentioning the location of the “odd” object (e.g., “The left side contains the cactus”). This was then followed by an affirmative or negative critical sentence. The affirmative version of the critical sentence began “It also contains the...” and the negative version began “It doesn’t contain the...”.

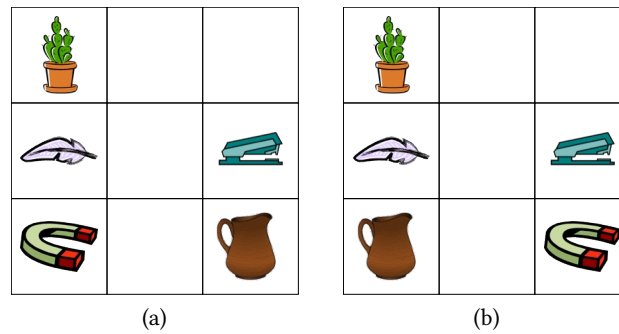


Figure 5.1: Example pair of eye-tracking experiment image stimuli. In the case of (a), the cactus is the object mentioned in the context sentence preceding the critical sentence (“The left side contains the cactus”). The target, appearing in the critical sentence, is either the magnet (in the affirmative case: “It also contains the *magnet*”) or the jug (pitcher; in the negative case: “It doesn’t contain the *pitcher*”). Thus, in the affirmative case, the possible candidates for the target during the critical window (just before the name of the target object) are the feather and the magnet; in the negative case, the candidates are the stapler and the pitcher. In the case of (b), the context sentence is identical, but the roles of *magnet* and *pitcher* are reversed in the sentences, and their locations in the grid are correspondingly reversed.

In each case the sentence concluded with the name of the target object. Only true sentences were presented. Thus, each pair of images was associated with four possible trials (two affirmative and two negative, with each target object featuring in one trial of each type). Only one of these trials was seen by any given participant.

Auditory stimuli were recorded by a female native speaker of American English and manipulated using the Audacity(R) recording and editing software (Audacity 2.1.2, Audacity Team, 1999–2016) and the SoX audio file manipulation utility (Chris Bagwell and SoX contributors, 2013) to splice the names of the requisite objects into standard recordings of the main parts of each sentence (“The left/right side / top/bottom row contains..” and “It also contains / doesn’t contain...”).

Each target item was also associated with a comprehension question which asked about either some detail of the object’s appearance (e.g. its colour) or its location in the grid, along the opposite axis to the location specified in the auditory input (e.g. “Was the magnet in the top row?”). The purpose of the comprehension question was to ensure that participants were incentivised to direct their attention as quickly as possible towards the target object, so as to take in as much information

as possible about this object before the grid disappeared.

In total, 10 pairs of grids were created, producing 480 possible trials. These were distributed across 8 versions of the experiment, each consisting of 60 trials. In addition to the experimental conditions, the described location (left column, right column, top row, or bottom row), the identity of the target object, and the locations of distractors were counterbalanced across participants.

5.2.2 Participants

Participants were 18 students from the University of Maryland community. Data from two participants were excluded from the analysis (one withdrew part-way through the experiment, and the second was a non-native English speaker). Of the 16 remaining participants, two were male, and the mean age was 21.3 years. All were native American English speakers (three were also bilingual). Participants provided written informed consent to participate and were compensated for their time with a payment of \$10.

5.2.3 Procedure

Participants were tested individually using an Eyelink 1000 eye-tracking apparatus, with an Eyelink 2000 camera (25 mm lens), tower mount, and forehead and chin rests to avoid head movements. At the start of the testing session each participant was seated approximately 1 m from a monitor where the visual stimuli would be displayed. Auditory stimuli were presented through speakers situated on either side of the monitor. The eye-tracking camera was trained on the participant's right eye, and the height and focus adjusted as required. The standard Eyelink nine-point calibration procedure, with fixation points displayed on the stimulus display monitor, was carried out at the start of the experiment to ensure that eye movements would be logged accurately. A maximum error of 0.5° was accepted; calibration was repeated if this threshold was exceeded. After successful calibration and validation, the experiment began. Calibration was repeated at the start of each block of trials, and between trials when excessive drift in the calibration was detected.

Trials were presented to the participant using the Experiment Builder software (SR Research), in the form shown in the example in Figure 5.2. A fixation target in the centre of the screen preceded each trial and the trial began when the Eyelink system recorded the participant as fixating on this target, in order to correct drift.

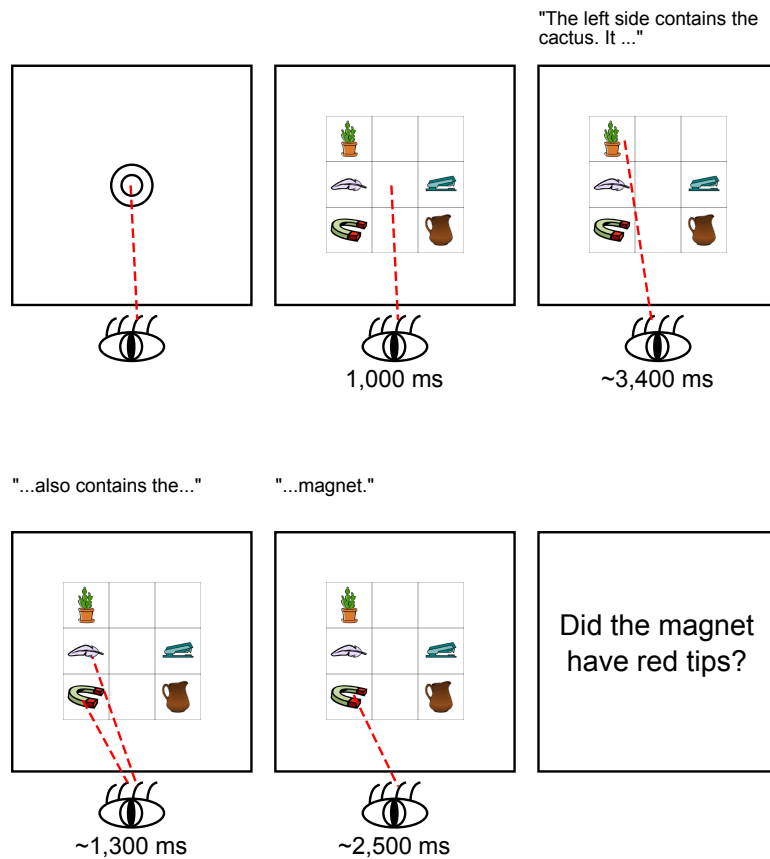


Figure 5.2: Eye-tracking experiment example trial with an affirmative sentence. To initiate the trial, the participant must fixate on a cue at the centre of the screen. The grid then appears and remains on the screen for 1,000 ms before the onset of the auditory stimulus. Dashed red lines represent regions where the participant might be expected to look for a disproportionate amount of time during each phase of the trial. After the grid disappears, a yes/no comprehension question related to the target object is displayed until the participant gives a response.

The visual stimulus (grid of images) then appeared on the screen. After 1,000 ms the auditory stimulus was played through the speakers. The visual stimulus remained on the screen during presentation of the context and critical sentence, and then for a further 1,000 ms (approximately 7,000 ms in total); participants'

monocular pupil locations were logged during the entire period, at a frequency of 2,000 Hz. After the image disappeared, a comprehension question asking about the target object was displayed on the screen and participants responded “yes” or “no” by pressing a button on a controller. Trials were presented in a random order in 6 blocks of 10 trials each, with the opportunity for participants to take a break (which was followed by recalibration of the eye-tracker) between blocks. The experiment lasted approximately 30 minutes.

5.3 Results

5.3.1 Comprehension response accuracy

Participants were readily able to answer the comprehension question accurately, with a mean response accuracy of 93.8% ($SD = 0.05$). The proportion of correct responses did not differ between affirmative ($M = 93.5\%$, $SD = 0.06$) and negative ($M = 94.0\%$, $SD = 0.04$) conditions, $t(15) = -0.31$, $p = .763$.

5.3.2 Eye movements

Data preparation

Fixation data were extracted for all trials except those with incorrect responses to the comprehension question (6% of all trials) and those with more than 2,000 missing fixations, e.g. because the camera had drifted out of calibration (5% of the remaining trials). Data from the remaining 853 trials were prepared for analysis as follows. First, the fixations from each trial were separated into three sets: those occurring before the critical time window of interest (before the onset of either *also* or *doesn't*); those occurring during this critical window (up to the onset of the name of the target object); and those occurring after the critical window (after the onset of the target). During the critical window, participants had enough information to deduce the row or side of the target object (as they knew it was either the same as that of the object mentioned in the critical sentence, in the case of hearing *also*, or the opposite, in the case of hearing *doesn't*), but not the specific identity of the target. Fixations were divided into these groups based on markers inserted into the eye movement logs at the precise onset of the relevant word in each trial.

Next, fixations within each window were allocated to time bins of equal length, each lasting approximately 100 ms. (The exact duration of each bin varied slightly

depending on the content of the particular trial, e.g. the name of the mentioned object in the context sentence.) Thirty-four bins were used for the pre-critical window, 13 for the critical window, and 25 for the post-critical window. Data sampled during saccades (as detected by the Eyelink system's automated processing) were discarded.

The location of each fixation was categorised based on its screen coordinates and the type of cell at that location on the trial in question. The categories were as follows: mentioned (the object mentioned in the context sentence preceding the critical sentence); target (the target object, i.e. the one mentioned in the critical sentence and the subject of the comprehension question); alternative to target (non-mentioned object in the same row or column as the target); foil (either of the two non-mentioned objects in the opposite row or column to the target); centre (the central cell, which was always blank, and where participants were always required to fixate at the start of the trial); other blank (any of the three non-central blank cells in the grid); and none (fixation outside the grid).

Within each time bin, the proportion of fixations to each type of cell was computed. The average proportion of fixations to each cell, across all participants, over the full time course of a trial is shown in Figure 5.3.

Analysis of fixations in the critical window

Before the critical window, participants began by fixating on the central, blank cell (they were required to fixate on a calibration point in this location in order to initiate the trial), and then shifted their attention to cells containing objects, fixating specifically on the one mentioned in the context sentence. After the critical window, the name of the target became available to participants, and they rapidly shifted their attention to the location of this object. These patterns are evident in Figure 5.3 for both affirmative and negative sentences. However, participants' fixations in the critical window between these two events, and how they differed between conditions, were of specific interest.

During this window, participants had already begun shifting their attention away from the object mentioned in the context sentence (because they knew that it would not be the target object), but did not yet know the identity of the target object. However, if they were able to incorporate incrementally all the information provided in the sentence so far, they could narrow down the possibilities for the target object based on its location relative to the one mentioned in the context sentence. Thus, the proportion of looks to the target and its alternative, vs. looks to

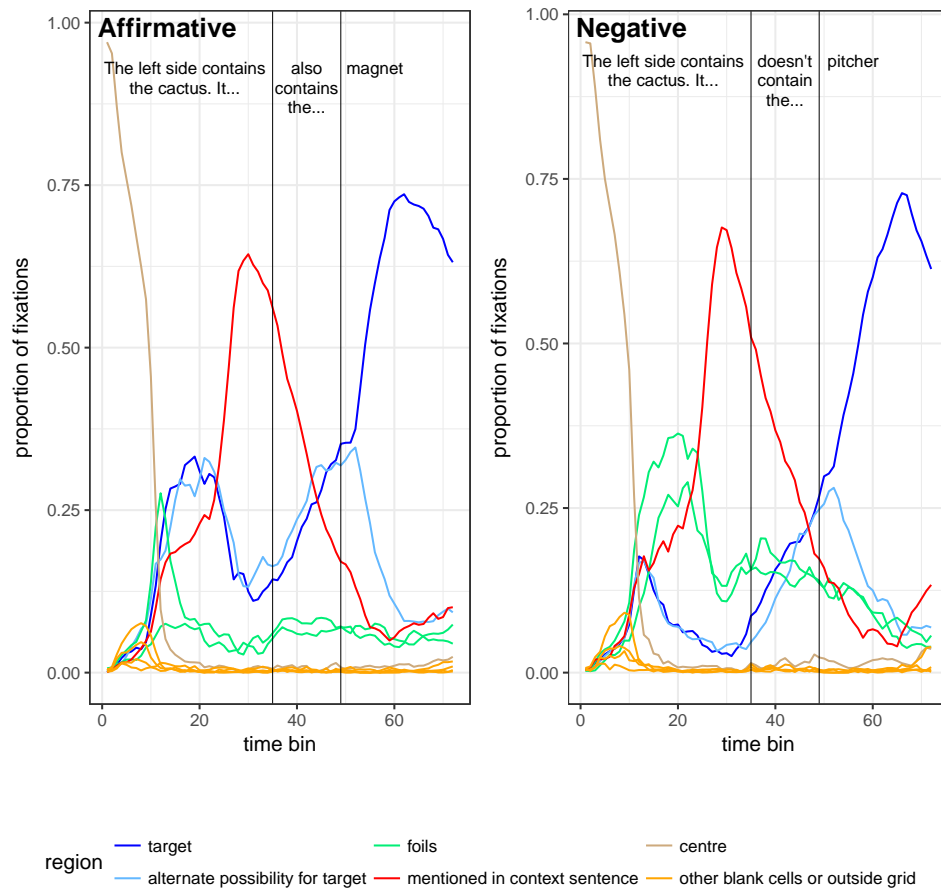


Figure 5.3: Eye-tracking experiment: the time course of eye movements during at average trial in each condition. Each time bin represents an average over approximately 100 ms. Note that the target and alternative (represented by blue lines) are located on the same side as the initially mentioned object in the affirmative case (left panel) and the opposite side in the negative case (right panel), whereas the opposite is the case for the foils (represented by green lines). Thus, the roles of these two pairs of objects in the early window are reversed between conditions.

either of the foils (the objects in the non-target location) could provide information on whether participants had taken into account all available information, including negation, in executing eye movements during this window.

Accordingly, a linear mixed regression model was constructed to investigate the proportion of fixations recorded to regions of interest during the critical window on each trial, using the `lmer` function of the R package `lme4` (Bates et al., 2015). Region type (target or alternative to target vs. foils) and sentence polarity

(affirmative or negative) were included in the full model as fixed factors, along with the interaction between these two factors. A random effect of polarity by participant was also included (more complex random effects structures caused the model to fail to converge). This full model was compared to models omitting the main effects and interactions of each factor in turn in order to establish which terms in the model represented significant effects.

The full model represented a better fit to the data than one including a fixed effect only of region type, $\chi^2(2) = 138.5$, $p < .001$, indicating that there was either a main effect of region type or an interaction involving this variable. The full model also represented a better fit to the data than one including a fixed effect only of polarity, $\chi^2(2) = 230.0$, $p < .001$, indicating that there was also either a main effect of polarity or an interaction involving this variable. Finally, the full model also represented a better fit to the data than a model including fixed effects of both these factors, but not the interaction between them, $\chi^2(1) = 138.5$, $p < .001$, indicating that there was a significant interaction between polarity and region type.

Coefficients representing the simple effects of each factor at each level of the other factor were computed based on the full model. These are listed in Table 5.1, and the overall pattern of effects is illustrated in Figure 5.4. Overall, these patterns show that negation had a detrimental effect on the proportion of fixations during the critical window to the possible candidates for the target object (i.e., the region containing the target and alternative), as opposed to the foils: in the affirmative condition, participants were more likely to look at the target or alternative (proportion of looks: $M = 0.47$, $SD = 0.35$) than at the foils ($M = 0.14$, $SD = 0.25$), whereas in the negative condition, they were approximately equally likely to look at the target or alternative ($M = 0.29$, $SD = 0.35$) and at the foils ($M = 0.32$, $SD = 0.35$).

Variable	β	t	d.f.	p	95% CI	
					lower	upper
Negation						
Target/target alternative	-0.18*	-8.48	1624.7	< .001	-0.22	-0.14
Foils	0.18*	8.49	1624.7	< .001	0.14	0.22
Foil region						
Affirmative polarity	-0.33*	-15.63	1679.1	< .001	-0.37	-0.29
Negative polarity	0.03	1.30	1679.1	.18	-0.01	0.07

Table 5.1: Eye-tracking experiment: simple effects, estimated for the full model, of each factor (polarity and region) at each level of the other factor. Reference values are affirmative for polarity and target/alternative for region. For example, the first line represents the difference in proportion of looks to the target and alternative in the negative condition, as compared to the affirmative condition.

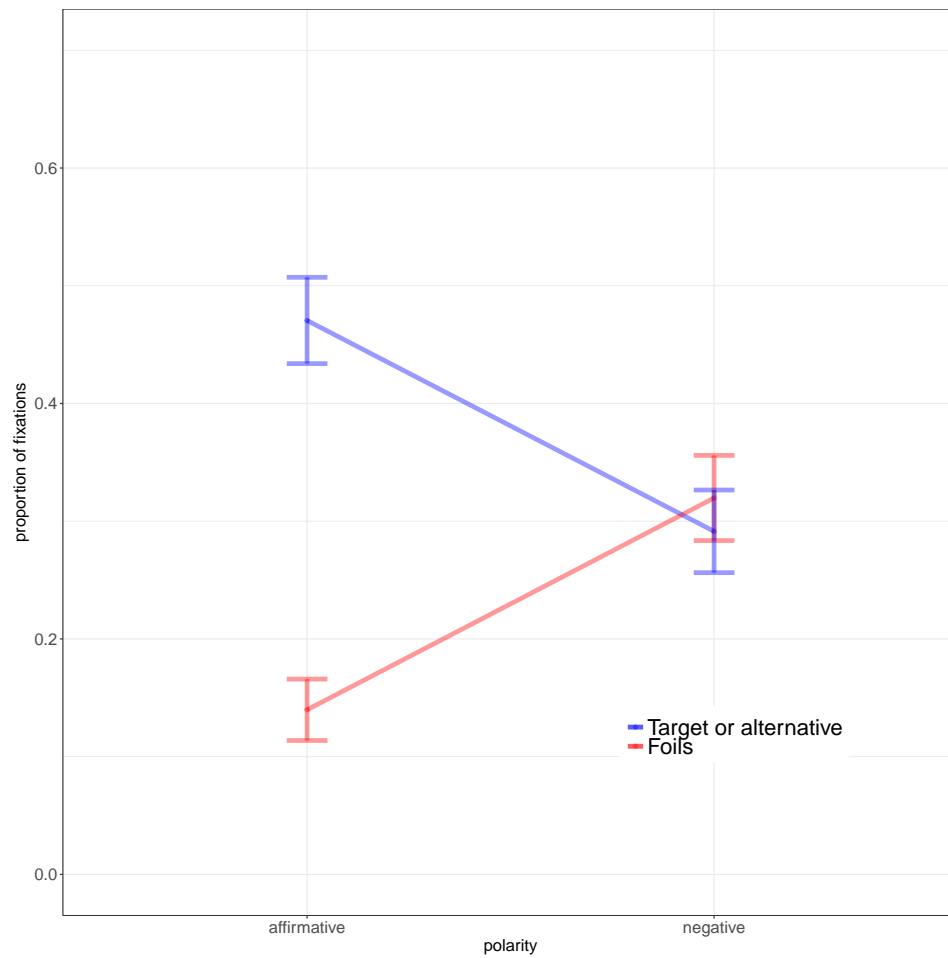


Figure 5.4: Eye-tracking experiment effects on proportion of looks: a summary illustration of the average proportion of looks, during the critical window, to each region on affirmative and negative trials. Error bars represent 95% confidence intervals.

5.4 Discussion

In this simple preliminary study looking at participants' anticipatory eye movements during processing of affirmative and negative sentences, participants were apparently better able to make use of all available information at an intermediate stage of input in the case of affirmatives. They were more likely to look at the possible candidates for the target object even before hearing the specific name of the target in the affirmative condition, while they were approximately equally likely to look at possible candidates and foils that could already be ruled out as targets in the negative condition.

However, this interpretation of the results ignores the fact that participants were more likely to already be looking at one of the candidates for the target in the affirmative condition. As shown in Figure 5.3, in the affirmative case, more than 25% of total fixations just before the onset of the word *also* were already on either the target or the alternative candidate for target, which were not yet designated as such; in contrast, at the same time point (just before *doesn't*) in the negative case, only about 12% of total fixations were already on either the target or the alternative candidate. This disparity in starting points arose as a result of the relationship between the positions of the object mentioned in the context sentence and the ultimate target object (and alternative candidate). On affirmative trials, the mentioned object was always in the same row or column as the target, and the opposite row or column to the foils.

After the onset of either *also* or *doesn't*, the proportion of looks to the possible candidates for the target then rose at approximately the same rate throughout the critical window until the target object was named explicitly; but because the starting point was lower in the negative condition, a comparison of the total proportion of fixations throughout the critical window showed that participants were more likely to look at the correct candidates in the affirmative case.

This aspect of the design had two separate confounding effects. First, the row or column containing the target was explicitly referred to during the context sentence (before the name of the mentioned object) on affirmative trials, whereas the row or column containing the foils was explicitly referred to on negative trials. Thus, participants' looks to the corresponding regions increased prior to the name of the mentioned object during the context sentence. Although a large proportion of looks (peaking at around 64%) were to the mentioned object following the onset of its name and before the onset of the critical word *doesn't* or *also*, this object was still

far from capturing all of participants' attention all of the time during this window, and most of the remaining fixations were to the objects in the same column or row. This is similar to an effect that was also in operation in the mouse-tracking experiments (Chapters 3 and 4), whereby the location initially mentioned was the relevant one in the case of affirmatives, and had to be subsequently suppressed in the case of negatives.

Second, when participants shifted their gaze away from the mentioned object during the critical window, they had further to travel to the candidates for the target in the case of negative trials (where they needed to cross the grid to the opposite column or row) than in the case of affirmative trials (where the candidates were located in the same column or row as the mentioned object). This increase in distance, although relatively small (effectively, the hypotenuse as opposed to the leg of an isosceles right triangle), may still have imposed costs: for example, the accuracy of saccades has an inversely proportional relationship with their amplitude (Kowler & Blaser, 1995). Furthermore, conceptually speaking, it is possible that the increased distance in the case of negative trials acted as a substantial barrier. Because the context sentence "carved up" the grid for a particular trial conceptually into either columns or rows, and the middle column or row (as applicable) was always empty, participants may have conceptualised the location of the mentioned object as having a disproportionately greater affinity with that of those objects in the same column or row, even if the physical distance was not much further.

This aspect of the design means that it is not easy to interpret the results of the experiment definitively. However, as it was intended primarily as a "proof of concept" or pilot study to investigate whether this type of visual world paradigm would work well as a way to index participants' anticipations about upcoming material with negative sentences in this type of episodic context, the concept was subsequently expanded to reduce the spatial confound and to investigate the impact of other factors.

To mitigate the problem in which affirmative sentences were always associated with a target in the same row or column as the object mentioned in the context sentence, and negative sentences with the reverse, two variations (affirmative and negative) on the context sentence were used in subsequent experiments, each of which could be paired with an affirmative or a negative critical sentence (see Table 5.2). While this design did not change the relationship between the location of the mentioned object and that of the target, it did dissociate the independent variable of interest from the particular column or row mentioned in the initial

Condition	Context sentence	Critical sentence
Aff-Aff	The left side contains the cactus...	...and it also contains the magnet
Aff-Neg	The left side contains the cactus...	...but it doesn't contain the pitcher
Neg-Aff	The right side doesn't contain the cactus...	...but it does contain the pitcher
Neg-Neg	The right side doesn't contain the cactus...	...and it doesn't contain the magnet

Table 5.2: Further eye-tracking experimental conditions: an expanded set of conditions used in subsequent experiments to mitigate the confounding effect of the region named in the context sentence.

context sentence.

In further experiments following a slightly distinct line of interest, the nature of the information available to participants about the objects in the grid was manipulated alongside the polarity of the critical sentence (and the context sentence, for counterbalancing, as described above). These designs reflected hypotheses, based on evidence that the shapes of objects and other related perceptual information are activated during language processing (e.g. Zwaan et al., 2002), that comprehenders need to be able to predict a specific full proposition (i.e., including the name of a possible target object) in order to correctly incorporate negation into that prediction and test it against the world (in this case, the visual world).

To investigate whether participants would produce anticipatory looks to the relevant part of the grid even when unable to make a specific prediction about the target object, an “invisible world” version of the experiment was conducted, in which the locations of all objects (except the one mentioned in the context sentence, which was displayed at all times) were indicated with crosses shown only briefly at the start of the grid presentation. Objects then appeared in full at the same time as the onset of the target word. In this case, participants’ performance in producing anticipatory looks to the target regions (during the equivalent critical window to the original experiment described here) was degraded for both affirmatives and negatives, but particularly in the latter case.

In a third experiment, to investigate whether partial information was sufficient to produce the same effect, silhouettes (present throughout the trial) were used in place of full objects, with the full image again revealed at the same time as the onset of the target object name in the sentence; in this case, results were more similar to the original experiment, with a smaller difference between affirmatives and negatives than in the “invisible world” case. These results are taken to tentatively support the hypothesis that information about the identity of potential candidates for a target (even if it is not accompanied by details of their appearance) is sufficient

to allow comprehenders to make and test predictions about upcoming material, including the incorporation of negation into these predictions. Analysis of the data and further work in this follow-up series of eye-tracking experiments is ongoing.

Overall, the preliminary eye-tracking experiment provided a somewhat promising avenue for further investigations in which predictability could be manipulated. Participants' anticipatory eye movements offer a particularly interpretable window into the time course of their prediction-making during sentence processing, which has not yet been exploited to its full potential in investigating the processing of negation.

Chapter 6

General Discussion

6.1 Summary of motivation

The work presented in this thesis was conducted to investigate factors affecting the incrementality of linguistic processing using a variety of methodological approaches probing online cognitive processes. The phenomenon of negation was of specific interest, as there is reason to believe this operation might present a special case in terms of incremental processing, arising both from its underlying nature and from previous research presenting an unclear picture on the extent to which negating elements can be incorporated incrementally by comprehenders into their partial representation of an utterance and used to update their predictions for upcoming material. Negation has been found to disrupt incrementality in some cases (Ferguson et al., 2008; Fischler et al., 1983; Lüdtke et al., 2008), but this disruption is reduced or disappears in other cases — primarily, when the discourse context supports the pragmatic use of negation (Dale & Duran, 2011; Johnson-Laird & Tridgell, 1972; Nieuwland & Kuperberg, 2008).

The specific research questions investigated here related to whether the predictability of later material in a sentence can influence the extent to which negation can be incorporated incrementally into the comprehender's interpretation of the partial sentence. Over the course of several experiments adopting different methodologies, the aim was to access participants' predictions about upcoming material as an index of the processing they were able to carry out online, while input was incoming. The focus was on manipulating predictability through the use of episodic scenarios rather than relying on participants' real-world knowledge to manipulate both the pragmatic felicity and the predictability of linguistic stimuli

in conjunction. The advantages of the use of episodic contexts for such manipulations include the avoidance of confounding factors arising from long-term semantic associations, and the possibility of presenting identical linguistic stimuli (accompanied by different contexts) across different conditions.

This approach was founded on the notion that predictability is an essential component of how an utterance can be understood as pragmatically felicitous to a greater or lesser degree. A felicitous utterance, in context, is informative, but achieves this without introducing information that has no reason to be brought into the discourse (Grice, 1975). Thus, the overarching hypothesis was that even when sentences were equally pragmatically felicitous, a reduction in the predictability of a critical word would disrupt processing to a greater extent in the case of negatives than in the case of affirmatives; under conditions of reduced predictability, comprehenders would be unable to update their predictions to account for the presence of negation and thus would initially make actively erroneous predictions, resolving this only at a later stage in processing.

6.2 Summary of findings

Chapter 2 presented an EEG-based experiment, in which participants viewed animations consisting of a unique object (on which one of two actions uniquely operated) and two other objects (on which the other action operated), then verified high- or low-predictability sentences describing the animation. The results of this experiment were both surprising and rather difficult to interpret theoretically. Although a reversal of the N400 responses to true and false sentences was observed in the case of negative sentences with low predictability, the same reversal was found in response to affirmative sentences with low predictability. In particular, the large amplitude of the N400 component in response to the critical word in a true affirmative sentence (especially one that was still relatively predictable, even in the low predictability condition) suggested that this paradigm did not capture exclusively the effects of participants' predictions for upcoming material. The most likely interpretation was that this pattern reflected the use of a strategy focusing on the "unique object" to ease working memory load, or perhaps automatic activation of the most salient (i.e., unique) object in the animation rather than the object most likely to complete the sentence. This approach was therefore found to be suboptimal to assess the research question.

In Chapter 3, a shift was made to a mouse-tracking methodology, with a

similar paradigm using static images (with specified locations in a grid) rather than animations. This enabled the expansion of the predictability factor to three levels without over-taxing participants' working memory and attention span. The use of three levels of predictability in turn enabled comparisons between levels (medium and low predictability) where neither stimulus contained a unique object, and the difference in salience between the targets may have been reduced. The results of this experiment were generally in line with the hypothesis (although a consistent main effect of negation, even in high-predictability conditions, was surprising), but the presence of the relevant interactions was somewhat inconsistent across different dependent variables. Importantly, the additional interaction of truth value with the independent variables of interest was problematic for the interpretation of the overall pattern.

As described in Chapter 4, the methodological pitfalls of the truth value judgement task were avoided in a further two closely-related mouse-tracking experiments, in which participants were tasked with selecting a picture to best complete a sentence rather than judging the truth value of a full sentence. This allowed more direct access to their predictions for upcoming material, and produced a surprising set of results, in which the opposite of the hypothesised pattern was observed: as well as a consistent main effect of negation once again, it was found that a reduction in predictability imposed a more detrimental effect on processing of affirmative than of negative sentences. This pattern was replicated across multiple dependent variables and obtained more clearly in mouse-tracking Experiment 3 (a speeded version of Experiment 2). The interpretation of its surprising nature was tackled in section 4.4, and is further discussed below (sections 6.3 and 6.4).

Finally, in Chapter 5, an alternative methodological avenue for investigating the effects of negation using episodic contexts was explored, namely eye-tracking. Predictability was not manipulated in this experiment, but the use of a visual world paradigm with linguistic stimuli describing the locations of objects in the visual world was found to show potential as a way of investigating processing differences between affirmative and negative sentences. Although clear differences in participants' eye movements were observed, with fewer looks during the critical window towards possible candidates for the eventual target in the negative condition, this could be attributed largely to the confounding effects in the exact nature of the task for negative as compared to affirmative sentences. Therefore, further work is necessary to explore how the incremental incorporation of negation into a sentence representation (or lack of such incorporation) affects comprehenders'

eye movements, particularly when the predictability of the target is manipulated.

6.3 Situating the findings relative to other work

Across all the experiments presented here, the findings have been (at least superficially) more in line with accounts of the processing of negative sentences that suggest that this can be more difficult than the processing of affirmatives even under very favourable conditions. This view is more consistent with non-incremental accounts of the processing of negative sentences than with accounts that suggest that negation can be strictly incrementally incorporated into partial sentence representations, although this evidence does not require a specific commitment to the view that the comprehender must wait for a full proposition to become available in order to carry out some form of two-stage processing. Additionally, even in this case, the findings do not necessarily distinguish between more specific accounts of the mechanism underlying non-incremental incorporation of negation, e.g. schema-plus-tag accounts (e.g. Mayo et al., 2004) vs multi-step simulation accounts (e.g. Kaup & Zwaan, 2003; Kaup et al., 2006). The firm statement that can be made about these findings is that they appear to be somewhat consistent with the view that negation is systematically more costly. Specifically, a clear distinction in processing between affirmatives and negatives was observed in every experiment except the first (Chapter 2), across almost all experimental conditions, suggesting that the latter incur some type of processing difficulty that is not inflicted by the former type. This was the case even though all experiments presented only sentences pragmatically licensed by the episodic context, a consideration that has been proposed as the essential factor in facilitating processing of negation. Under this interpretation, it is necessary to conclude that in cases where equivalence between affirmatives and negatives is observed (e.g. Nieuwland & Kuperberg, 2008), this can be attributed to special circumstances elicited by presenting sentences in a pragmatic context specifically designed to enhance the felicity of negative sentences.

However, in the work presented here, consistent interactions between predictability and polarity were also observed in the experiments with the most methodologically reliable and statistically robust findings (Chapter 4). Therefore, an interpretation in which negation simply requires an extra processing step, even when its use is pragmatically licensed (but not pragmatically supported by a rich context), is inadequate to fully explain the results. High levels of predictability

appear to facilitate processing of affirmative sentences to a greater extent than they facilitate processing of negative sentences. This is discussed further below.

First, however, it is important to situate the type of stimuli presented in this work in the context of those used in similar research. A comparison of both the pragmatic felicity and the predictability of linguistic input is necessary to explore possible reasons for the differences between these and other findings. In the experiments described here, a deliberate choice was made to manipulate predictability through the use of episodically generated contexts, rather than using real-world knowledge or associations. As discussed in the relevant chapters, this approach has some advantages (such as the ability to control predictability and other aspects of the stimuli precisely), but it also introduced certain trade-offs. In particular, the artificial or constrained nature of the episodic scenarios may have meant that the pragmatics of the full situation comprising each stimulus were atypical or somewhat lacking in context, despite the careful licensing of the sentences by the context. This characterisation applies to a greater extent to the static grids and descriptions used in the experiments described in Chapters 3, 4 and 5; the animations presented in the EEG experiment (Chapter 2) provided a richer contextual grounding for the accompanying sentences, and more specific pragmatic reason to deny the occurrence of events (actions that were attempted but failed to take effect). However, the use of animations presented its own challenges: each trial was lengthy, placing limits on the amount of data that could be collected, and it would be difficult to expand this paradigm to scenarios with multiple levels of predictability, as the task was already taxing for participants. Thus, across most of the experiments presented here, the sentences were probably best characterised as pragmatically licensed, but not contextually enriched. This distinction is probably one with a rather fuzzy boundary, but might be illustrated by comparing the sentences presented by Dale and Duran (2011) in their Experiment 2 (“You want to lift an elephant? Elephants are not small”) vs. their Experiment 3 (“‘You want to lift an elephant?’ the mother said to her child, ‘but elephants are not small.’”) In the former case, the simple preamble means that the negation does not emerge out of the blue. However, the latter case goes beyond this to present a miniature discourse in which the underlying communicative intent for the entire utterance is clear.

The sentences presented across the present experiments were in at least one sense very strongly predictable compared to those used in other experiments, and to sentences usually encountered in natural dialogue. Given the constrained contexts,

which amounted to miniature “worlds” with small, fixed sets of populating entities, the high predictability sentences only had one referent that could complete them truthfully, and even the conditions with lower predictability presented only two or three options for the critical word. In experiments presenting a mixture of true and false sentences (Chapter 2’s EEG experiment and Chapter 3’s mouse-tracking Experiment 1), participants could not logically make a guaranteed prediction for which word would in fact complete the sentence, even when only one candidate would make it true, as a false sentence was equally likely. However, as rapid prediction-making (as an integral component of sentence processing) tends to be automatic rather than strategic, the ratio of false to true sentences encountered in the experimental setting did not necessarily affect this. In sentence completion (Chapter 4; mouse-tracking Experiments 2 and 3) and eye-tracking experiments (Chapter 5), participants never encountered any false sentences, meaning that any potential effect of the true:false ratio was not in operation.

In this sense, then, the stimuli were predictable to an artificially high degree. However, as noted above, they may in some cases have been presented in a context that uses an atypically artificial means of pragmatic licensing, and furthermore, the predictions that participants could make were based on immediately preceding, episodic information. While comprehenders are certainly able to take such information into account in their assessments of these sentences, there is a sense in which this type of prediction may be more difficult for comprehenders to generate online, compared to the type of predictions that must be generated during natural language comprehension. Sentences presented within a natural discourse with a rich contextual embedding, striking a pragmatic balance between relevance and informativity, may in some ways be easier to predict because comprehenders can base their hypotheses on their understanding of the speaker’s intent and their real-world knowledge. These predictions may be easier to make because they are facilitated by strong, long-term associations between concepts. For example, although the sentence “A robin is a...” could be accurately completed by a number of possible expressions (*bird*, *creature*, *symbol of Christmas*, *delightful chap*, and so on), the semantic associations it presents may make these predictions easier to generate than a prediction that relies on a temporary and asemantic association between the entity and an arbitrary predicate (e.g. “the top row contains”). Furthermore, the surrounding discourse context, in a natural language setting, may provide rather extensive clues as to which of those predictions might be best (an encyclopedia article vs. a poem, for instance). In contrast, somewhat arbitrary

associations that are relevant only episodically have none of these advantages and additionally impose a working memory burden on comprehenders (in storing these associations in short-term memory) before prediction-making can even be considered. The ease of episodically calculated predictions and more long-term association-based predictions could be important in the case of negations, in light of the following discussion.

6.4 Global and local processing

It is integral to the role of negation that operations involving negation activate concepts or entities specifically for the purpose of suppressing them in favour of their opposite (to the extent that this is defined). For example, in a sentence of the type used in Chapters 3 and 4 such as “The top row doesn’t contain the ...”, the concept of the *top row* and objects located there are initially activated, and the comprehender must use the fact that *doesn’t* subsequently appears to transfer their attention from objects associated with the top row to those associated with its “opposite” (which, in this case, is clear to them through a combination of general world knowledge and the specific context of the experimental scenario: the bottom row). Thus, there is a crucial distinction to be made in the processing of negation between what is activated locally (for example, at earlier points of an utterance before a negating element occurs, or before its implications for the full representation of the sentence have been fully computed) and what needs to be activated by the globally-interpreted representation of the sentence following the incorporation of negation.

The burden imposed on processing by this feature of negation (at least in the compositional form of a negating particle and verb or other predicate), and the extent to which it impedes incremental updating and prediction-making during input, depends on several factors. Simple word order is important: if negation appears before the negated concept (as in the type of sentences used in Chapter 2: “The wizard didn’t raise the position of the basket”), it may be possible for the comprehender to adjust their response to this concept commensurately. Conversely, when negation appears after the concept to be negated (as in the type used in Chapter 3: “The top row doesn’t contain the basket”), there may be further effects at work in which a greater distance between the two may impose an increasing level of difficulty on suppressing what has already been strongly activated. (Such activation arises not necessarily only from single words; it can

be based on compositional analysis of preceding parts of the sentence. Studies manipulating the predictability of words using real-world knowledge often demonstrate or indeed rely on this: for example, the specific prediction *palms* is made as a completion for “They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of...” as a result not of any single word, but thanks to contributions to the full meaning of the sentence made by *tropical*, *resort*, and *planted*; Federmeier & Kutas, 1999.) However, other factors also appear to matter for the incremental processing of negative sentences, including those at issue in the work here: contextual pragmatics and predictability. These are probably important because of the way in which they govern the ease with which the “nuisance” activation arising from the negated concept can be replaced with the most relevant alternative.

In particular, contextual pragmatics and predictability dictate the extent to which a clear contrast frame for the negated concept is available. As discussed in the General Introduction and in Chapter 4, predicates may have a clear opposite (e.g., *open* vs. *closed*; see Kaup et al., 2006; Mayo et al., 2004), in which case a negation has a ready-made, obvious meaning. Alternatively, there may be multiple alternatives (as in the “multary” contexts presented by Orenes et al., 2014), in which case the comprehender is more likely to retain activation of the negated concept as the only possible way or representing the set of all likely meanings. In the case of an intermediate stage during online processing of a partial proposition, in which negation has only just appeared, the extent to which the comprehender can make accurate predictions depends on the potential contrast sets narrowed down by earlier parts of the sentence. For instance, in the case of “A robin is not a...” (Fischler et al., 1983), the scope of possible dimensions along which negation could occur is almost infinite. Critically, the semantic associations of the negated concept are not the only factor contributing to the extent to which this computation is possible or easy: contextual factors also contribute. In contrast to the “robin” sentence, in the case of a more pragmatically felicitous utterance embedded in a richly specified context, the speaker’s intent in uttering a negation is likely to be relatively clear to the comprehender (for instance, as in the enriched sentences presented by Dale & Duran, 2011). This information, combined with an incrementally updated representation of the critical part of the utterance, allows the comprehender to formulate an accurate prediction about upcoming material that incorporates the meaning of the negation.

In the sentences presented in the mouse-tracking experiments discussed here,

two factors relating to predictability seem to be important in modulating the ease with which accurate predictions based on negation can be made. First, the experimental context that applied across all trials, in which the same type of grid was consistently presented and only objects actually present in the grid were ever named (Chapter 3) or presented as response options (Chapter 4), made it clear to participants that the expression *the top row doesn't* could be taken as having a meaning or predictive capacity exactly equivalent to that of *the bottom row*, and vice versa. (If the experiments had included sentences in which objects not present in the grid were mentioned, a possible completion to “The top row doesn't contain...” could always have been a non-present object; in fact, because this was not the case, possible completions were always objects located in the bottom row.) This information about the “world” in which the sentences were presented meant that, within the context, the linguistic stimuli were pragmatically licensed and their meaning picked out in a similar way to the facilitation of more “natural” sentences based on realistic linguistic context. Second, the predictability of the sentences once all the information was incorporated, including negation, was governed by the manipulation of the visual context. Thus, on critical negative trials, participants needed to switch their attention from the row containing three objects to a row containing variably one, two, or three objects, depending on condition. This switching process and the associated predictions to be made were probably influenced by the variation in salience of rows with different numbers of items, as well as the presence of negation in itself, as shown by the polarity \times predictability interaction described in Chapter 4 (section 4.4.3).

In both these cases, relationships constructed artificially by the experimental design, rather than through deeply embedded semantic associations, underlaid the computations necessary to make predictions. As discussed above, this choice about the experimental paradigms employed was made for well-founded reasons, but it may contribute to how the results should be interpreted in light of what predictability means in this context. Negating the predicate *open* to arrive at *closed*, in a pragmatic context lending clear support to this interpretation of the speaker's intent, may be a very different process requiring different strategies to the operations involved in negating the predicate *top row* and its arbitrary, episodically-generated associations with three everyday objects to arrive at *bottom row* and its similar associations with two different and semantically unrelated objects, even when the experimental context makes this process of operations very clear analytically.

In summary, while negation often induces the need for a switch in the subject of attention or activation, and predictability may be one of the factors governing how easy or difficult this switch is, multiple facets of predictability may influence the difficulty in different ways, affecting the extent to which comprehenders can make accurate predictions accounting for the presence of negation.

6.5 Limitations and generalisability

The experiments described here present a mixed picture of the impact of varying levels of predictability on the incremental interpretation of negation. However, the most methodologically sound finding was replicated (Chapter 4) and several potentially interesting avenues explored. The generalisability of these results is restricted to some degree by many of the factors discussed above: it is not necessarily clear how the findings generalise to cases of variable predictability in settings that are not governed by the type of clear and constrained context generated by the experimental paradigms, for example, nor even whether it is meaningful to refer to this type of predictability in a natural discourse context, in which predictability of upcoming material is rarely dictated by the existence of a clear set of candidates. However, the interpretations proposed above do clarify the nature of the relationship between contextual predictability governed by pragmatic considerations, and the type of analytical or non-semantic predictability manipulated here. The difference between the two may represent an important factor in the incremental interpretation of negation.

Some further limitations that apply across these experiments include the fact that it remains unclear how to interpret the mouse-tracking findings in detail without postulating that a floor effect was involved, even though the data support the view that there was not such an effect present; it would therefore be useful to understand in more detail the patterns that emerge in terms of trajectory clustering across a wider range of paradigms. The experiments were also restricted, as discussed above, to exploring a small part of the predictability spectrum, with even the low predictability conditions involving relatively high cloze probabilities. It would be difficult to explore sentences with lower predictability still while maintaining the reliance on episodic contexts for manipulation, rather than shifting to a paradigm relying on long-term knowledge.

Because different types of sentences were used across experiments employing different methodologies, it is not always clear exactly what caused the differences

in findings across experiments. This means that not all effects obtained here have been accounted for; some remain difficult to understand.

Finally, this series of experiments does not provide evidence to support or contradict specific theories of the mechanisms underlying processing of negation, even in cases where it suggests that a two-step mechanism might be in play under certain circumstances; for example, this evidence cannot distinguish between tagging and simulation theories.

6.6 Future work

Two lines of further investigation have been extended from the work presented here. First, as described briefly in Chapter 5, the eye-tracking paradigm piloted in the initial experiment reported on has been developed to investigate further the impact of providing varying amounts of information on potential targets to the participants. This series of experiments reflects the hypothesis that, at least under some circumstances, negation cannot be incorporated fully into a comprehender's interpretation of a sentence until a complete proposition is available for negation, and thus that the ability to predict a fully fleshed-out candidate proposition (including, in this case, the specific name of a target object) governs the ease with which the information from a negating element can be included. This is tested through the use of images that make only location information or only a silhouette initially available to the participant.

Second, a report on a Russian-language replication of the picture choice mouse-tracking experiments (Chapter 4) is in preparation. Russian's somewhat freer word order allows the negating element (*net*) to be placed in either a late or an early position in the sentence relative to the predicate that is negated (glosses: "This time is / is not on the top row..." vs. "This time on the top row is / is not..."). This manipulation allows for an exploration of the extent to which activation of the location containing the foil objects before the negation is heard is the driver of the effects observed in the English-language version (which is equivalent to the late-position Russian condition).

Future work could also use the eye-tracking paradigm to examine the effects of manipulating predictability in a similar way to the mouse-tracking paradigm. Participants' eye movements in these cases could provide further and clearer evidence on the extent to which a predictability manipulation affects their ability to make predictions on the basis of negation, as well as a clearer window on the

time course with which they are able to do so. For experiments of this type, it is important to account for the fact that confounds with polarity can operate as a result of the fact that attention may be drawn to a particular part of the visual world prior to the critical part of the linguistic input; this should be avoided or accounted for in analyses.

Finally, it would be interesting to examine the effects of presenting sentences of the type used in this work with lower predictability than those used in the experiments here. Although this presents a methodological challenge (because it would be difficult to simply expand the paradigm further without entirely overloading participants' working memory and preventing them from completing the task), a solution would allow for more direct comparison between results in this type of paradigm and those from research such as Nieuwland and Kuperberg (2008) and Dale and Duran (2011), in which the sentences presented are highly predictable as well as pragmatically felicitous, but in most cases still not as absolutely predictable as the ones used here.

6.7 Conclusion

Overall, this series of experiments highlights a set of circumstances in which negative sentences incur a processing cost, relative to affirmatives, even when all sentences are pragmatically licensed and in fact highly predictable. Contrary to the original hypotheses, this relative cost actually diminished rather than increasing upon reducing the critical word's predictability: that is, varying predictability in this paradigm had a larger effect of affirmatives than on negatives, suggesting that it is relatively difficult for comprehenders to formulate accurate predictions during this kind of negative sentence at all, even when there are few candidates for the prediction. This can probably be attributed to the clash between local associations and the global meaning of a partial sentence representation that occurs in the case of many negative sentences, especially when (as here) they are not supported by a richly-constructed and therefore highly constraining pragmatic context.

Bibliography

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419–439. doi: 10.1006/jmla.1997.2558.
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1.
- Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33, 583–609. doi: 10.1111/j.1551-6709.2009.01022.x.
- Alxatib, S., & Pelletier, F. J. (2011). The psychology of vagueness: Borderline cases and contradictions. *Mind and Language*, 26(3), 287–326. doi: 10.1111/j.1468-0017.2011.01419.x.
- Anderson, S. E., Huette, S., Matlock, T., & Spivey, M. J. (2010). On the temporal dynamics of negated perceptual simulations. In F. Parrill, V. Tobin, & M. Turner (Eds.), *Meaning, Form and Body* (pp. 1–20). Stanford, CA: CSLI Publications.
- Apostel, L. (1972a). Negation. *Logique et Analyse*, 15(57–58), 209–317.
- Apostel, L. (1972b). The relation between negation in linguistics, logic and psychology: A provisional conclusion. *Logique et Analyse*, 15(57–58), 333–401.
- Atlas, J. D. (1977). Negation, ambiguity, and presupposition. *Linguistics and Philosophy*, 1(3), 321–336.
- Audacity Team. (1999–2016). *Audacity 2.1.2*. URL: <https://www.audacityteam.org>.
- Ayer, A. J. (1952). Negation. *Journal of Philosophy*, 49(26), 797–815.
- Barres, P. E., & Johnson-Laird, P. N. (2003). On imagining what is true (and what is false). *Thinking and Reasoning*, 9(1), 1–42. doi: 10.1080/13546780244000097.

- Barsalou, L. W. (2010). Grounded cognition: Past, present and future. *Topics in Cognitive Science*, 2, 716–724. doi: 10.1111/j.1756-8765.2010.01115.x.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01.
- Berg, P., & Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and Clinical Neurophysiology*, 90, 229–241. doi: 10.1016/0013-4694(94)90094-9.
- Berto, F. (2015). A modality called ‘Negation’. *Mind*, 124(495), 761–793. doi: 10.1093/mind/fzv026.
- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391–416.
- Chomsky, N. (1970). Deep structure, surface structure and semantic interpretation. In R. Jakobson & S. Kawamoto (Eds.), *Studies in General and Oriental Linguistics* (pp. 183–216). Tokyo: TEC Corporation for Language Research.
- Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2016). A ‘bag-of-arguments’ mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31(5), 577–596. doi: 10.1080/23273798.2015.1066832.
- Coles, M. G. H., & Rugg, M. D. (1995). Event-related brain potentials: An introduction. In M. G. H. Coles & M. D. Rugg (Eds.), *Electrophysiology of mind: Event-related brain potentials and cognition*. Oxford: Oxford University Press.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31(3), 187–276. doi: 10.1016/0010-0277(89)90023-1.
- Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, 35, 983–996. doi: 10.1111/j.1551-6709.2010.01164.x.
- Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, 35, 15–28. doi: 10.3758/BF03195938.
- De Mey, M. (1972). The psychology of negation and attention. *Logique et Analyse*, 15(57–58), 137–153.
- DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8, 1117–1121. doi: 10.1038/nn1504.
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: long-term memory

- structure and sentence processing. *Journal of Memory and Language*, 41, 469–495. doi: 10.1006/jmla.1999.2660.
- Ferguson, H. J., Sanford, A. J., & Leuthold, H. (2008). Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research*, 1236, 113–125. doi: 10.1016/j.brainres.2008.07.099.
- Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1), 11–15. doi: 10.1111/1467-8721.00158.
- Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W., Jr. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, 20(4), 400–409. doi: 10.1111/j.1469-8986.1983.tb00920.x.
- Fischler, I., Childers, D. G., Achariyapaopan, T., & Perry, N. W., Jr. (1985). Brain potentials during sentence verification: Automatic aspects of comprehension. *Biological Psychology*, 21, 83–105. doi: 10.1016/0301-0511(85)90008-0.
- Freeman, J. B., & Ambady, N. (2010). Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, 1, 226–241. doi: 10.3758/BRM.42.1.226.
- Frege, G. (1960). Negation. In P. Geach & M. Black (Eds.), *Translations from the Philosophical Writings of Gottlob Frege* (pp. 67–137). Oxford: Oxford University Press.
- Geach, P. (1965). Assertion. *The Philosophical Review*, 74(4), 449–465.
- Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59(4), 601–613. doi: 10.1037/0022-3514.59.4.601.
- Giora, R. (2006). Anything negatives can do affirmatives can do just as well, except for some metaphors. *Journal of Pragmatics*, 38, 981–1014. doi: 10.1016/j.pragma.2005.12.006.
- Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C. (1999). Not propositions. *Journal of Cognitive Systems Research*, 1, 19–33. doi: 10.1016/S1389-0417(99)00004-2.
- Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, 4(2), 107–111.
- Greenberg, J. H. (1966). *Language Universals*. Mouton: The Hague.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics (vol. 3)* (pp. 41–58). New York: Academic Press.
- Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical*

- Transactions of the Royal Society*, 362, 801–811. doi: 10.1098/rstb.2007.2089.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistics Society, Series C (Applied Statistics)*, 28(1), 100–108. doi: 10.2307/2346830.
- Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation? An examination of negated metaphors. *Journal of Pragmatics*, 38, 1015–1032. doi: 10.1016/j.pragma.2005.12.005.
- Horn, L. R. (1989). *A natural history of negation*. Chicago: The University of Chicago Press.
- Huang, Y. T., & Snedeker, J. (2011). *Logic and conversation* revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, 26(8), 1161–1172. doi: 10.1080/01690965.2010.508641.
- Huette, S., & Anderson, S. (2012). Negation without symbols: The importance of recurrence and context in linguistic negation. *Journal of Integrative Neuroscience*, 11(3), 295. doi: 10.1142/S0219635212500239.
- Humberstone, L. (2000). The revival of rejective negation. *Journal of Philosophical Logic*, 29(4), 331–381. doi: 10.1023/A:1004747920321.
- Jespersen, O. (1917). *Negation in English and other languages*. Copenhagen: Høst.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, Massachusetts: Harvard University Press.
- Johnson-Laird, P. N., & Tridgell, J. M. (1972). When negation is easier than affirmation. *Quarterly Journal of Experimental Psychology*, 24, 87–91.
- Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, 10, 244–253.
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 29, 133–156. doi: 10.1016/S0749-496X(03)00023-8.
- Kaup, B. (2001). Negation and its impact on the accessibility of text information. *Memory and Cognition*, 29(7), 960–967. doi: 10.3758/BF03195758.
- Kaup, B., & Lüdtke, J. (2007). The experiential view of language comprehension: How is negation represented? In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 255–288). London: Lawrence Erlbaum Associates.
- Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with

- contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38, 1033–1050. doi: 10.1016/j.pragma.2005.09.012.
- Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 29(3), 439–446. doi: 10.1037/0278-7393.29.3.439.
- Kazanina, N. (2017). Predicting complex syntactic structure in real time: Processing of negative sentences in Russian. *Quarterly Journal of Experimental Psychology*, 70, 2200–2218. doi: 10.1080/17470218.2016.1228684.
- Kent, C., Guest, D., Adelman, J. S., & Lamberts, K. (2014). Stochastic accumulation of feature information in perception and memory. *Frontiers in Psychology*, 5, 412. doi: 10.3389/fpsyg.2014.00412.
- Kent, C., Taylor, S., Taylor, N., & Darley, E. (2017). Tracking trajectories: A brief review for researchers. *The Cognitive Psychology Bulletin*, 2(1), 12–16.
- Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24(5), 541–599. doi: 10.1080/20445911.2012.660913.
- Kim, A. E., Oines, L. D., & Sikos, L. (2016). Prediction during sentence comprehension is more than a sum of lexical associations: the role of event knowledge. *Language, Cognition and Neuroscience*, 31(5), 597–601. doi: 10.1080/23273798.2015.1102950.
- Kowler, E., & Blaser, E. (1995). The accuracy and precision of saccades to small and large targets. *Vision Research*, 35(12), 1741–1754. doi: 10.1016/0042-6989(94)00255-K.
- Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science*, 4(12), 463–470. doi: 10.1016/S1364-6613(00)01560-6.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 14.1–14.27. doi: 10.1146/annurev.psych.093008.131123.
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207, 203–205. doi: 10.1126/science.7350657.
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307, 161–163. doi: 10.1038/307161ao.

- Lea, R. B., & Mulligan, E. J. (2002). The effect of negation on deductive inferences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28(2), 303–317. doi: 10.1037//0278-7393.28.2.303.
- Löbner, S. (2000). Polarity in natural language: Predication, quantification and negation in particular and characterizing sentences. *Linguistics and Philosophy*, 23(3), 213–308. doi: 10.1023/A:1005571202592.
- Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.
- Lüdtke, J., Friedrich, C. K., de Filippis, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence-picture verification paradigm. *Journal of Cognitive Neuroscience*, 20(8), 1355–1370. doi: 10.1162/jocn.2008.20093.
- MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15(4), 633–642.
- Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, 39(2), 89–105. doi: 10.1016/0010-0277(91)90039-7.
- Mayo, R., Schul, Y., & Burnstein, E. (2004). ‘I am not guilty’ vs ‘I am innocent’: Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, 40, 433–449. doi: 10.1016/j.jesp.2003.07.008.
- Mayo, R., Schul, Y., & Rosenthal, M. (2014). If you negate, you may forget: Negated repetitions impair memory compared with affirmative repetitions. *Journal of Experimental Psychology*, 143(4), 1541–1552. doi: 10.1037/a0036122.
- McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19, 22–24. doi: 10.1111/j.1467-9280.2008.02041.x.
- NaturalSoft Limited. (2015). *NaturalReader version 11*. URL: <https://www.naturalreaders.com>.
- Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42.
- Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An Event-Related Potential study on the pragmatics of negation. *Psychological Science*, 19(12), 1213–1218. doi: 10.1111/j.1467-9280.2008.02226.x.
- Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... Huettig, F. (2018). Large-scale replication study reveals a limit on

probabilistic prediction in language comprehension. *eLife*, 7, e33468. doi: 10.7554/eLife.33468.

- Nieuwland, M. S., & van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7), 1098–1111. doi: 10.1162/jocn.2006.18.7.1098.
- Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, 77, 25–39. doi: 10.1016/j.jml.2014.08.002.
- Nygaard, L. C., & Pisoni, D. B. (1995). Speech perception: New directions in research and theory. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition: Speech, language, and communication* (pp. 63–96). San Diego: Academic Press. doi: 10.1016/B978-012497770-9.50005-4.
- Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, 74, 36–45. doi: 10.1016/j.jml.2014.04.001.
- Orenes, I., Moxey, L., Scheepers, C., & Santamaría, C. (2016). Negation in context: Evidence from the visual world paradigm. *The Quarterly Journal of Experimental Psychology*, 69(6), 1082–1092. doi: 10.1080/17470218.2015.1063675.
- Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3), 105–110. doi: 10.1016/j.tics.2006.12.002.
- Price, H. (1990). Why 'not'? *Mind*, 99(394), 221–238.
- Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, 80, 4–9. doi: 10.1016/j.biopsycho.2008.05.002.
- Ripley, D. W. (2011). Negation, denial, and rejection. *Philosophy Compass*, 6, 622–629.
- Russell, B. (1905). On denoting. *Mind*, 14(56), 479–493.
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109–147. doi: 10.1016/S0010-0277(99)00025-6.
- Shramko, Y. (2005). Dual intuitionistic logic and a variety of negations: The logic of scientific research. *Studia Logica*, 80(2/3), 347–267.
- Sturt, P., & Lombardo, V. (2005). Processing coordinated structures: Incrementality and connectedness. *Cognitive Science*, 29, 291–305. doi: 10.1207/s15516709cog0008.

- Sullivan, M. W. (1967). *Apuleian Logic*. Amsterdam: North-Holland.
- Tian, Y., & Breheny, R. (2015). Dynamic pragmatic view of negation processing. In P. Larrivée & C. Lee (Eds.), *Negation and polarity: Experimental perspectives*. Switzerland: Springer. doi: 10.1007/978-3-319-17464-8_2.
- Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology*, 63(12), 2305–2312. doi: 10.1080/17470218.2010.525712.
- Tomlinson, J., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, 69, 18–35. doi: 10.1016/j.jml.2013.02.003.
- Urbach, T. P., DeLong, K. A., & Kutas, M. (2015). Quantifiers are incrementally interpreted in context, more than less. *Journal of Memory and Language*, 83, 79–96. doi: 10.1016/j.jml.2015.03.010.
- Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, 63, 158–179. doi: 10.1016/j.jml.2010.03.008.
- van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research*, 3, 701–718. doi: 10.1016/S0926-6410(03)00196-4.
- Wansing, H. (2016). Falsification, natural deduction and bi-intuitionistic logic. *Journal of Logic and Computation*, 26(1), 425–450. doi: 10.1093/logcom/ext035.
- Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, 11, 92–107.
- Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, 52, 133–142.
- Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, 4(1), 7–11.
- Wason, P. C. (1972). In real life negatives are false. *Logique et Analyse*, 15(57–58), 17–38.
- Wason, P. C., & Jones, S. (1963). Negatives: Denotation and connotation. *British Journal of Psychology*, 54, 299–307.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. <http://ggplot2.org>.
- Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (2013). Validating the truth of propositions: behavioral and ERP indicators of truth

evaluation processes. *Social Cognitive and Affective Neuroscience*, 8, 647–653. doi: 10.1093/scan/nss042.

Wode, H. (1977). Four early stages in the development of L1 negation. *Journal of Child Language*, 4, 87–102. doi: 10.1017/S0305000900000490.

Xiang, M., Grove, J., & Giannakidou, A. (2016). Semantic and pragmatic processes in the comprehension of negation: An event related potential study of negative polarity sensitivity. *Journal of Neurolinguistics*, 38, 71–88. doi: 10.1016/j.jneuroling.2015.11.001.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, 13, 168–171. doi: 10.1111/1467-9280.00430.